

S

S²PD – Surface Severe Plastic Deformation

► [Surface Nanocrystallization and Hardening \(SNH\)](#)

Safflower Oil

► [Natural Oils as Lubricants](#)

Saliva Lubrication

JASON R. STOKES

School of Chemical Engineering, The University of Queensland, Brisbane, QLD, Australia

Synonyms

[Oral lubrication](#); [Saliva tribology](#)

Definition

Saliva lubrication concerns the tribological properties between two interacting surfaces that are, or have been, in contact with saliva. This includes the tribological properties of a substrate that has been in contact with saliva *ex vivo* (i.e., expectorated) or *in situ* (i.e., in mouth) for a period of time so that an adsorbed proteinaceous film has formed on one or both surfaces. Saliva lubrication, as considered here, only includes circumstances when the adsorbed film is hydrated.

Scientific Fundamentals

Lubrication is a primary function of saliva and many other mucous biofluids. Lubrication from saliva is provided through the presence of both a mobile and adsorbed salivary film coating on surfaces in the mouth. The adsorbed film coating is a renewable multicomponent proteinaceous biofilm on teeth and mucosal surfaces (the

“acquired pellicle”). The mobile saliva film is the source of the proteins in the pellicle and also serves to keep the pellicle hydrated, which is necessary for its lubricating function. Lubrication is required for protecting oral surfaces from damage, wear, and irritation during teeth contact, speech (tongue movements), mastication, and swallowing, as well as to facilitate the movement of food.

About Saliva

Saliva is produced from three major glands around the oral cavity: the parotid, submandibular, and sublingual. Parotid saliva is secreted through paired ducts situated in the cheek, while submandibular/sublingual saliva is secreted through ducts under the tongue surface. There are also numerous minor glands situated throughout the oral cavity. The saliva produced from each of the glands vary in composition and physical properties (e.g., rheology), and they are also not well mixed in the mouth. For whole mouth saliva (WMS) secreted during the resting state (i.e., unstimulated), parotid saliva contributes 25%, submandibular 60%, and sublingual 7–8% to the volume of WMS, and the overall flow rate is ~0.3–0.4 mL/min, although these amounts vary between and within individuals. All are rich in proteins, but parotid saliva has a Newtonian rheology with a viscosity equal to that of water, while submandibular/sublingual saliva is more viscous, shear thinning, and highly elastic (Stokes and Davies 2007). Mechanical chewing and acidic solutions are two primary methods to stimulate the production of saliva to flow rates exceeding 1 mL/min. While saliva is approximately 98% water, it is a highly complex mixture of proteins, super-saturated minerals, debris, and bacteria. There have been as many as 309 proteins detected in whole saliva and 130 in the acquired enamel pellicle (Dawes 2008). The major proteins are acidic and basic proline-rich proteins, low and high molecular weight glycoproteins (Muc5b and Muc7), agglutinins, cystatins, histatins, statherin, and amylase. Many of these proteins are surface active and have anti-viral and anti-bacterial properties. Also present are electrolytes that keep saliva hypotonic so there is a tendency for water to be absorbed across the oral mucosa. The composition, flow rate, and physical properties of saliva, both as a whole and from

each of the major glands, is dependent on the method of stimulation and a large number of other factors such as time of day, body position, previous stimulation, circannual rhythms, and drug use. Due to variability in saliva samples obtained between individuals and within the same individual, extreme care is needed to ensure consistent sampling methodologies including ensuring donors have not eaten for at least 1–2 h prior to collection of expectorated saliva for ex vivo testing. Saliva's properties can also rapidly change within minutes of expectoration, centrifugation, and exposure to air, as well as following immediate freezing at -80°C . To limit changes over short periods of time, saliva samples should be stored under ice in sample vials that resist protein adsorption.

Lubricating Function

Saliva serves numerous and multiple functions in the oral cavity, including solubilization of food, bolus formation, food and bacterial clearance, dilution and cleansing, digestion of starch and lipids, mineralization, buffering of acids, and protection against erosion and dental caries. However, one of its key attributes is formation of a lubricating film to protect tooth surfaces against abrasion and attrition, and to reduce trauma to soft tissues during mastication, swallowing, and speaking. Abrasion occurs due to the rubbing of foreign bodies against teeth surfaces, including that which occurs from the use of abrasive toothpastes. Attrition occurs due to repeated contact between opposing teeth. The acquired pellicle lubricates teeth surfaces to protect it from such frictional wear, and this lubricating film is also renewable due to rapid adsorption of salivary proteins to any exposed enamel. Lubrication of the soft tissue is provided by the presence of a protein-rich thin film coating of saliva that also assists in keeping oral surfaces from drying out. A loss of salivary film coating on soft tissue leads to more susceptibility to abrasion and irritation.

A variety of measurements on the lubricating properties of saliva and its components have been made using a variety of surfaces, including enamel/teeth surfaces, mucosal surface, and orthodontic materials. Friction coefficients range from 0.004 to 0.45 for a variety of sliding speeds, loads, and substrates. An in vivo tribometer has also been utilized to examine the friction of steel and Teflon surfaces sliding against oral surfaces directly (Olsson et al. 1991) to assess the efficacy of saliva substitutes for treating xerostomia. However, most studies have used unsophisticated tribological equipment with a limited range of loads and speeds, while the roughness and wetting characteristics of the surfaces used are

typically not quantified, which makes it difficult to evaluate and compare saliva's lubricating properties between techniques. The following sections concentrate on studies that focus on the lubrication properties of saliva itself using modern tribology equipment and substrates with controlled surface chemistry and roughness. Also highlighted is the surface film arising from adsorption of proteins in saliva that is responsible for its boundary lubricating properties.

Wear of In Situ Acquired Pellicle on Teeth Surface

The formation of an acquired salivary pellicle on teeth surfaces results from the adsorption of biopolymers at the tooth-saliva interface. Salivary proteins adsorb almost instantaneously onto the enamel surface and an initial pellicle film is formed after only a couple of minutes. A second protein adsorption phase occurs for 30–90 min until there is a plateau in the thickness of the pellicle. To study the tribological properties of an in situ formed pellicle, it is necessary to mount specimens in devices worn in the mouth for a period of time and remove them to perform ex vivo measurements. It is found that an acquired pellicle formed in situ significantly reduces enamel and dentine wear from toothpaste abrasion. In one example, the wear arising from brushing an enamel surface using a particular toothpaste is reduced significantly from 0.23 to only 0.03 microns due to the presence of an in situ formed salivary pellicle (Joiner et al. 2008). However, friction measurements on in situ formed saliva pellicles are yet to be conducted.

Formation of Acquired Pellicle from Ex Vivo Saliva

Ex vivo saliva readily adsorbs and wets both hydrophilic and hydrophobic substrates. Adsorption studies have been carried out on model substrates using ellipsometry, quartz crystal microbalance with dissipation monitoring (QCM-D), and surface plasmon resonance (SPR). These studies have been used to understand the adsorbed film thickness, hydration, and viscoelasticity of the acquired pellicle and how it interacts with oral care and food formulations. Studies usually involve exposure of substrates to expectorated whole saliva, which is typically diluted to $\sim 10\%$ in a physiological buffer, while monitoring the formation of the adsorbed salivary film. Two recent in vitro studies using silica surfaces (Cardenas et al. 2007) and polydimethylsiloxane (PDMS)-coated surface (Macakova et al. 2010) show that the adsorbed salivary film is structurally heterogeneous with a uniform, thin, and dense inner layer formed by small salivary proteins and

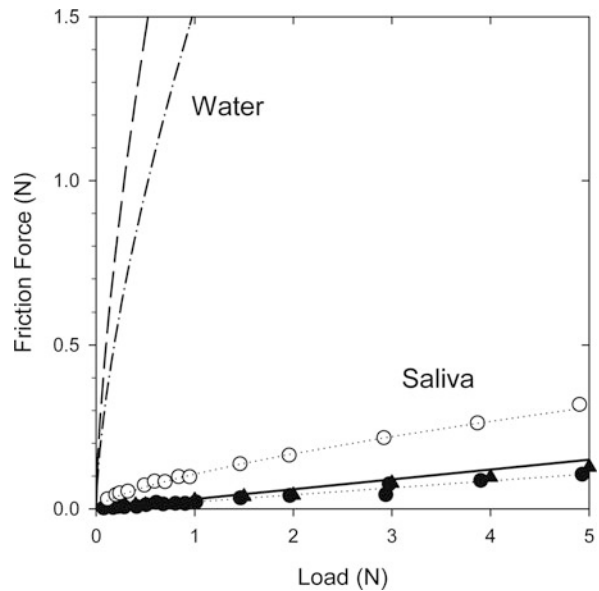
non-glycosylated parts of salivary mucins, and a thicker outer layer that is mucin-rich. The inner layer is likely to be a monolayer of various globular proteins in saliva with a film thickness of ~ 4 nm, while the overall adsorbed film thickness was around 20–30 nm (Cardenas et al. 2007). The acquired pellicle is a highly viscoelastic solid-like film on hydrophobic surfaces, with the inner layer being relatively rigid and strongly adhered (anchored) to the surface, while the upper layer consists of extended biopolymers that are highly hydrated ($>80\%$ water). The viscoelasticity of the adsorbed salivary film is highly responsive to its environment. For example, decreases in ionic strength lead to swelling of the outer layer, which is typical for adsorbed polyelectrolyte polymers and thus considered to be due to mucin glycoproteins that are multivalent amphiphilic polyelectrolytes. The film thickness following exposure to ion-free water rose to 44 nm but subsequently collapsed to 6 nm; the collapse is thought to be driven by electrostatic attraction between different parts of the multicomponent film (Macakova et al. 2010).

Saliva Lubrication: In vitro Measurements from Ex Vivo Saliva

A range of techniques have been used to examine the lubricating properties of saliva, as stated previously, and collectively these show that saliva lowers the friction coefficient (μ) for most surfaces in comparison to water alone. Recent studies are considered here that study saliva lubrication using atomic force microscopy (AFM) and a modified tribometer.

Berg et al. 2003 performed nano-tribological measurements on saliva films formed through ex vivo adsorption of 10% WHS (unstimulated) using an AFM. The lateral friction force was measured while sliding a silica bead against hydrophilic silica substrates under loads of up to 300 nN. The presence of the freshly formed saliva film reduced the friction coefficient between hard silica surfaces by a factor of 20 (Hertzian contact pressure, P_H ca. 100 MPa), with a linear relationship found between friction force (F) and load (w) so that $\mu = 0.03$. This contact pressure is characteristic of occlusal contacts (i.e., contact between opposing teeth).

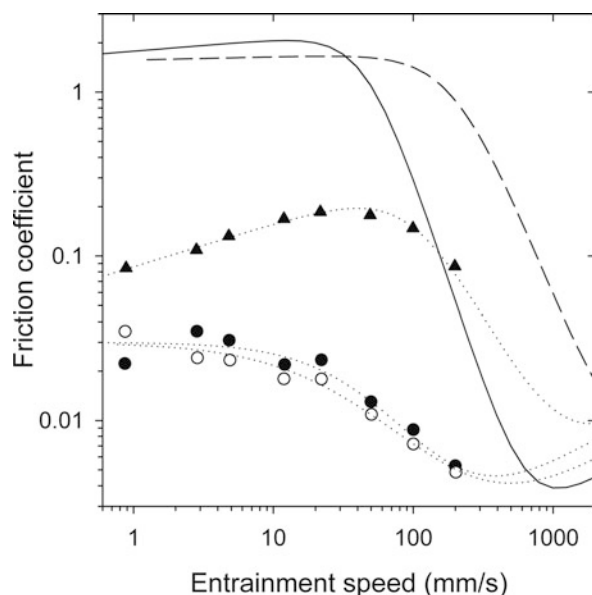
Bongaerts et al. (2007a) and Macakova et al. (2011) used a modified tribometer to characterize saliva's lubricating properties in a rolling/sliding contact (slide-to-roll ratio, $S = 50\%$) with a PDMS ball and disk to mimic the low modulus and wetting characteristics of the soft mucosal surfaces that are intrinsically hydrophobic when not coated by a saliva film. The low modulus of the substrate also means that the contact pressure (P_H ca. 0.1 MPa) is low enough that the pressure does not influence the



Saliva Lubrication, Fig. 1 Friction force versus load in the boundary regime for saliva between PDMS ball-disk at $S = 50\%$ and $U = 5$ mm/s for mechanically stimulated WMS samples from (Bongaerts et al. 2007) ((● smooth, $F = 0.02 W$; ○ rough, $F = 0.11 W^{2/3}$) and (Macakova et al. 2011) (▲, smooth) at 35°C . Also shown are the friction force for water under the same conditions for smooth (—, $F = 2.3 W^{2/3}$) and rough PDMS (---, $F = 1.5 W^{2/3}$), and line for $F = 0.03 W$ obtained for 10% saliva sample between silica bead/surface on AFM (Berg et al. 2003) (—) based on data at loads below 300 nN

rheological properties of the lubricant (i.e., it is in the iso-viscous elastohydrodynamic domain). This contact pressure is characteristic of pressure arising from interaction with a soft tissue or soft food material. Tribology measurements are conducted immediately upon expectoration of WMS, and salivary proteins are found to rapidly adsorb to the surfaces to create a very effective boundary film. At an entrainment speed of $U = 5$ mm/s, the system is in the boundary lubrication regime for smooth (9 nm r.m.s. roughness) and rough (380 nm r.m.s. roughness) PDMS surfaces.

Figure 1 shows the friction force in the boundary regime as a function of load for hydrophobic PDMS lubricated with mechanically stimulated WMS. For smooth PDMS, $F = 0.02 w$ for the two WMS samples from (Bongaerts et al. 2007a; Macakova et al. 2011). This relationship is remarkably consistent with the result of Berg et al. (2003), despite this being on smooth hard hydrophilic silica surfaces and under nano-tribological conditions. For rough PDMS substrates lubricated with



Saliva Lubrication, Fig. 2 Stribeck friction curve for mechanically stimulated saliva between hydrophobic PDMS ball/disk with $S = 50\%$ at 35°C . Shown are averaged friction data from (Bongaerts et al. 2007) for fresh (\bullet) and centrifuged (\circ) WMS for smooth PDMS, and fresh WMS for rough PDMS (\blacktriangle). The *solid* and *dashed* lines are master curves for aqueous fluids for smooth and rough PDMS, respectively, based on water's viscosity (data from (Bongaerts et al. 2007)). *Dotted lines* are anticipated Stribeck curves for saliva based on the measured data and master curve (using saliva viscosity at $\dot{\gamma} = 1,000 \text{ s}^{-1}$)

a WMS sample, higher friction coefficients are observed and $F = 0.11w^{2/3}$. The $w^{2/3}$ scaling is expected for highly compliant surfaces where the area of contact is determined by elastic conformity rather than plastic deformation of asperities, and it is also observed for water as the lubricant for smooth and rough PDMS surfaces. This scaling is not observed for saliva on smooth surfaces because the adsorbed biopolymer film thickness ($\sim 20\text{--}30 \text{ nm}$) is greater than the surface roughness so that it carries the load in the contact.

The dependence of the friction coefficient on entrainment speed of WMS lubricant is shown in Fig. 2 for PDMS surfaces. WMS is effective in the boundary and mixed lubrication regimes due to the presence of the adsorbed film, but the hydrodynamic regime could not be accessed due to the low viscosity (η) of saliva, which is only marginally greater than that of water (Stokes and Davies 2007). At low entrainment speeds ($U < 20 \text{ mm/s}$) the boundary friction coefficient for the saliva films shown

was $\mu \sim 0.02$ and $\mu \sim 0.1\text{--}0.2$ for smooth and rough PDMS respectively, while $\mu \sim 2$ for water. The friction coefficient for the smooth surface is similar to that found for saliva on other smooth surfaces (Berg et al. 2003; Gans et al. 1990), while the value for rough surface is close to those observed for ex vivo saliva films on rough mucosal surface (porcine tongue) (Prinz et al. 2007) and enamel (Aguirre et al. 1989). The friction coefficient decreases further with increasing entrainment speed as it enters the mixed lubrication regime.

The friction coefficient for water is shown in Fig. 2 in the form of a so-called master curve. A master curve is obtained by plotting μ versus ηU at a constant load using a range of Newtonian aqueous fluids (including water) that do not contain any species that adsorb to either surface. The master curve from Bongaerts et al. (2007b) is re-plotted in Fig. 1 by dividing ηU by the η of water at 35°C . Since the hydrodynamic regime for aqueous fluid master curves are the same regardless of whether the surfaces are hydrophobic or hydrophilic, it is anticipated that the friction in the hydrodynamic regime for saliva is primarily governed by its viscosity at the high shear rates present in the contact, although the fluid's elastic properties may also play a role as discussed below. Thus, the dotted lines in Fig. 2 have been drawn using the equations for the master curve modified to fit the friction coefficients for saliva in the boundary and mixed lubrication regime and using the viscosity of saliva at shear rate $\dot{\gamma} = 1,000 \text{ s}^{-1}$ in the equations for the master curves. This approach suggests that under the right conditions a minimum friction coefficient of 0.004 would be obtained at the junction between the mixed and hydrodynamic regime.

The physical properties of saliva, including its rheology, alter dramatically during storage time and upon centrifugation. For example, fresh WMS is highly viscoelastic while WMS that is stored, aged, and/or centrifuged is inelastic although their viscosity remains largely unchanged. This loss of elasticity can be observed visually by no longer being able to form a long standing thread when saliva is stretched between thumb and forefinger. However, despite this dramatic change, there is little difference in the lubricating properties between centrifuged and fresh saliva, as shown in Fig. 2, while aged saliva exhibited slightly higher friction coefficients (Bongaerts et al. 2007). This means that the viscoelasticity of saliva does not play a role in its boundary lubricating properties. It also means that samples can be prepared so that cells and other debris can be removed from the sample without affecting its boundary lubricating properties.

The lubricating properties of saliva are inherently linked to the adsorption of salivary proteins to the

contacting surfaces. A large proportion of the decrease in friction coefficient can be assigned to the hydrophilic nature of the adsorbed saliva film, although the films structure and viscoelasticity is considered very important for wear resistances and general robustness. For example, its robustness is exemplified by the film retaining its lubricating properties on smooth PDMS surfaces upon sequential drying and rehydration during rubbing (Bongaerts et al. 2007a). The saliva film is also able to maintain its ability to lubricate up to the maximum testing load of 5 N. However, when the ex vivo formed saliva film is rinsed so that only the lubrication of the adsorbed layer is examined under physiological salt conditions (70 mM NaCl), the friction coefficient increases at loads above 2 N that may be due to shear-induced wear of the film. The effect becomes more extensive as the ionic strength of the solvent is lowered below physiological conditions (Macakova et al. 2011). Therefore, having excess saliva beyond what is physically adsorbed may assist in protecting the surfaces from wear and maintaining a low boundary friction coefficient, potentially through re-adsorption of salivary proteins in areas where the adsorbed film is compromised.

In Situ Versus Ex Vivo Pellicle

At this stage it is unknown how the nature of the ex vivo formed salivary film differs from an in situ or in vivo formed pellicle. In situ formed pellicles are created over much longer time scales and are much denser and thicker than in ex vivo experiments, and clearly the wear studies on toothpaste brushing abrasion demonstrate the robustness of in situ formed pellicles. In addition, some studies have shown that transglutaminase, which is generated in the oral cavity by buccal epithelial cells, is able to form covalent crosslinks between many salivary proteins and may also cross-link proteins in the pellicle to form a stronger film than what can occur ex vivo. Thus, further developments are still needed to improve ex vivo saliva pellicles as a mimetic for in situ or in vivo formed pellicles.

Saliva Variability and Viscoelasticity

A failing of the published literature on saliva lubrication and adsorbed film properties is that it only includes data from a few individuals. Given that the properties and composition of saliva varies between and within individual donors, the exact values obtained from only a few donors are only an indication of saliva's properties and should be treated with some caution until further extensive studies with a large number of donors is conducted. However, the author has recently completed a study using freshly expectorated WMS samples from a larger number of donors ($n = 16$) following mechanical and acid

stimulation and characterized using the method and smooth PDMS surfaces of Bongaerts et al. (2007). Average friction coefficients of $\mu = 0.021$ (standard error, s.e. 0.003) and $\mu = 0.0098$ (s.e. 0.002) are found for mechanical and acid stimulated WMS, respectively. While acid-stimulated WMS is considerably more viscoelastic than mechanically stimulated saliva, no statistical link is found between the boundary friction coefficient and WMS viscoelasticity. In addition, the viscosity of acid and mechanically stimulated saliva is not substantially different (Stokes and Davies 2007).

Many review articles on saliva infer that there is a link between its lubrication function and its viscoelasticity. However, there is no evidence whatsoever that these are related, as highlighted, but it is still worthy of consideration. Essentially, measurements on saliva lubrication have only been made in the boundary and mixed regime. The boundary friction coefficient is intricately linked to the *adsorbed* salivary film properties, and although the adsorbed film may be viscoelastic, this is not associated with saliva's elastic properties in bulk fluid flow. While the hydrodynamic regime is yet to be evaluated, it is hypothesized that it is in this regime that saliva's elastic properties may play an important role (Stokes and Davies 2007).

Compared to a standard polymer solution, saliva has very unusual rheological characteristics; it shows extremely high elasticity for a very low viscosity fluid (Stokes and Davies 2007). For WMS saliva derived from acid stimulation, the primary normal stress difference (N_1) that is measured in a cone-and-plate rheometer from the normal force generated under shear, is 100 times that of the shear stress (σ) across a broad shear rate range. In comparison, even the relatively inelastic mechanical stimulated saliva had a stress ratio of $N_1/\sigma \sim 10$. It should be noted that Newtonian fluids are inelastic such that $N_1 = 0$, while values of $1 < N_1/\sigma < 10$ are typically found for polymer solutions that are considered to be highly elastic. The remarkably high stress ratios for saliva indicate that elastic stresses dominate over viscous stresses during flow, and they are considered to arise from the presence of extremely high molecular weight glycoprotein (mucin) that forms super-molecules by aggregating end-to-end. Following a simple analysis of second order fluid between two dimensional sliding thrust bearing of length L , the extra lift force predicted to arise from normal stresses is of order $(N_1/\sigma)(h/L)$, where h is the film thickness. Hence, in circumstances where $L/h < 100$, saliva's normal stresses are anticipated to provide additional support of the applied load. This may be particularly important in the movement of food around the oral cavity where contact forces may be considerably

lower, and film thicknesses between food and oral surfaces higher, than those occurring during direct interaction between oral surfaces.

Key Applications

Oral Health

Saliva maintains oral health. Lubrication in the oral cavity is provided by formation of an acquired pellicle and from the presence of a mobile thin film coating. The importance of this function is exemplified by the medical condition of dry mouth, referred to as xerostomia, which is common to hyposalivators. Hyposalivation (<0.1 mL/min WMS flow rate) is a common side effect of many therapeutic drugs, radiation treatment for head and neck cancer, and Sjorgren syndrome. A low flow rate of saliva leads to a localized decrease in salivary film thickness (e.g., from 100 to 10 μm on the hard palate), which is considered to impair oral lubrication (Dawes 2008). Xerostomia leads to difficulties in swallowing, speech, and mastication, as well as altered taste sensations and a general deterioration in oral health including increased abrasion and dental caries at teeth surfaces. This can also have a serious and detrimental effect on well being and cause considerable anxiety and stress on patients. Understanding the lubrication properties of saliva is being used to assist in the development and evaluation of effective “artificial” salivas and other strategies to alleviate the effects of xerostomia.

Oral Care

Toothpaste and toothbrushing, as well as mouth rinses, are designed to improve oral health both through cleansing and removal of adhered bacteria (in the acquired pellicle, which is a biofilm) and various debris from foods and beverages. However, there is a balance between removal of the salivary pellicle from teeth surfaces and wear of the enamel underneath (Joiner et al. 2008). Understanding the lubrication function of the pellicle, and ways in which to measure and manipulate its properties, will ultimately allow the improved design of oral care products to aide in the long-term protection of oral surfaces from frictional wear.

Oral Processing of Food

The intraoral salivary coating interacts with dietary components, which consequently influences both taste and tactile perception (mouthfeel). Studies on the influences of foods on saliva lubrication are being used to improve knowledge of oral processing and determination of how surface friction and food-saliva interactions influence sensory perception and oral deposition/depletion processes.

For example, the polyphenols in teas, wines, and other beverages containing phytonutrients aggregate with salivary proteins. This aggregation causes depletion of the salivary film from oral surfaces, which leads to a loss of oral lubrication and a dry mouth sensation that alters the sensory properties of the beverage or food (Rossetti et al. 2009). In addition, food and beverages can induce different levels of saliva production that can subsequently affect their organoleptic properties. For example, water induces very low amounts of relatively inelastic saliva to be produced, which can give the sensation of a dry mouth and increased oral friction. However, acidic beverages generate large volumes of highly elastic saliva, but the acid erodes the pellicle from teeth surfaces to give a “furry” or rough sensation (Davies et al. 2009). The stimulation of saliva in response to acid is considered to occur from increased production of submandibular saliva to renew the pellicle on teeth surfaces. But while the response of saliva and the pellicle to food and beverages has implications towards the health and protection of oral surfaces, it also has implications on the mouthfeel and other sensory percepts during and following eating and drinking. Thus, improving our understanding of saliva lubrication and its measurement may lead to improved design of healthier food and beverages with consumer acceptable sensory properties.

Cross-References

- Oral Tribology
- The Tribology of Dental Materials
- Tribology of Foods
- Tribology of Teeth Cleaning

References

- A. Aguirre, B. Mendoza, M.J. Levine, M.N. Hatton, W.H. Douglas, In vitro characterization of human salivary lubrication. *Arch. Oral Biol.* **34**, 675–677 (1989)
- I.C.H. Berg, M.W. Rutland, T. Arnebrant, Lubricating properties of the initial salivary pellicle – an AFM Study. *Biofouling* **19**, 365–369 (2003)
- J.H.H. Bongaerts, D. Rossetti, J.R. Stokes, The lubricating properties of human whole saliva. *Tribol. Lett.* **27**, 277–287 (2007a)
- J.H.H. Bongaerts, K. Fourtouni, J.R. Stokes, Soft-tribology: Lubrication in a compliant PDMS-PDMS contact. *Tribol. Int.* **40**, 1531–1542 (2007b)
- M. Cardenas, U. Elofsson, L. Lindh, Salivary mucin MUC5B could be an important component of in vitro pellicles of human saliva: An in situ ellipsometry and atomic force microscopy study. *Biomacromolecules* **8**, 1149–1156 (2007)
- G.A. Davies, E. Wantling, J.R. Stokes, The influence of beverages on the stimulation and viscoelasticity of saliva: Relationship to mouthfeel? *Food Hydrocoll.* **23**, 2261–2269 (2009)
- C. Dawes, Salivary flow patterns and the health of hard and soft oral tissues. *J. Am. Dent. Assoc.* **139**, 18S–24S (2008)

- R.F. Gans, G.E. Watson, L.A. Tabak, A new assessment in vitro of human salivary lubrication using a compliant substrate. *Arch. Oral Biol.* **35**, 487–492 (1990)
- A. Joiner, A. Schwarz, C.J. Philpotts, T.F. Cox, K. Huber, M. Hannig, The protective nature of pellicle towards toothpaste abrasion on enamel and dentine. *J. Dent.* **36**, 360–368 (2008)
- L. Macakova, G.E. Yakubov, M.A. Plunkett, J.R. Stokes, Influence of ionic strength changes on the structure of pre-adsorbed salivary films. A response of a natural multi-component layer. *Colloids Surf. B Biointerfaces* **77**, 31–39 (2010)
- L. Macakova, G.E. Yakubov, M.A. Plunkett, J.R. Stokes, Influence of ionic strength on the tribological properties of pre-adsorbed salivary films. *Tribol. Int.* **44**, 956 (2011)
- H. Olsson, V. Henriksson, T. Axell, A new device for measuring oral mucosal surface friction – reference values. *Scand. J. Dent. Res.* **99**, 329–332 (1991)
- J.F. Prinz, R.A. de Wijk, L. Huntjens, Load dependency of the coefficient of friction of oral mucosa. *Food Hydrocoll.* **21**, 402–408 (2007)
- D. Rossetti, J.H.H. Bongaerts, E. Wantling, J.R. Stokes, A.M. Williamson, Astringency of tea catechins: more than an oral lubrication tactile percept. *Food Hydrocoll.* **23**, 1984–1992 (2009)
- J.R. Stokes, G.A. Davies, Viscoelasticity of human whole saliva collected after acid and mechanical stimulation. *Biorheology* **44**, 141–160 (2007)

Saliva Tribology

► Saliva Lubrication

Salt Bath Nitriding

► Tufftriding and Tennifer Surface Treatment

Scanning Electron Microscopy (SEM)

CHAOYING NI

Department of Materials Science & Engineering,
University of Delaware, Newark, DE, USA

Definition

Scanning electron microscopy is a characterization technique that images and analyzes a specimen by scanning an accelerated electron beam, followed by selectively collecting and recording secondary electrons, back-scattered electrons, and other signals arising from the

beam and specimen interactions. A modern scanning electron microscope (SEM) permits the observation and characterization of materials at micro- to nano-scale.

Scientific Fundamentals

Major Components

Figure 1 provides a schematic diagram of an SEM, including the electron source (also called electron gun or electron emitter), condenser lenses, scan coils, and detectors. A high vacuum is required to operate an SEM. A brief description for each of the major components is presented below.

Electron Gun

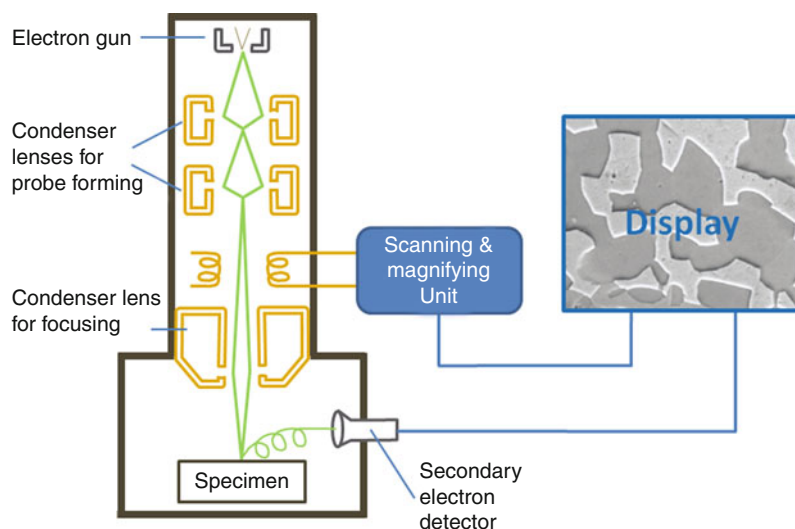
Similar to the conventional light microscope, an electron microscope has a source of illumination. However, in contrast to the conventional light microscope that employs visible light, the electron microscope utilizes an electron beam emitting from an electron source. Based on the principles of electron emission, there are generally two types of electron sources: thermionic emission and field emission.

The emitter of a thermionic source is usually made of tungsten (W) hairpin or lanthanum hexaboride (LaB_6). Electron emission occurs as individual electrons reach an energy level to overcome a barrier called work function (ϕ), measured in electron volt (eV). For thermionic emission, the electron energy required for emission (ϕ) is provided by applying a filament current. The work function of tungsten is about 4.7 eV, while that of LaB_6 is about 2.6 eV.

The emitter of a field emission source consists of usually an oriented single-crystal W needle coated with ZrO_2 to reduce work function. The field emission source is also called a field emission gun (FEG). There are two types of FEG: cold FEG and Schottky FEG.

Cold FEG emits electrons solely based on a phenomenon called electron field emission, which can be explained by a theory of electron quantum tunneling. According to this theory, when the W tip in ultra-high vacuum experiences a strong static electric field, electrons in the W tunnel out from the tip as if there were no resistance or barrier. In FEG, the static electric field is established by applying an extracting voltage beneath the W needle. The electric field at the W tip is inversely proportional to its radius and typically reaches more than 1,000 V per micron. Cold FEG works at room temperature.

In Schottky FEG, electron emission takes place in a so-called field-and-temperature regime, where both the thermal energy from filament current and a static electric field



Scanning Electron Microscopy (SEM), Fig. 1 Schematic diagram of SEM, including major components

contribute to electron emission. This phenomenon can be understood as thermionic emission occurring under the enhancement of static electric field, which effectively reduces the work function.

Two major parameters characterizing electron sources are emission current density and brightness. Emission current density is expressed by the Richardson equation $J = AT^2 e^{-\frac{\phi}{kT}}$, where J is emission current density (SI unit: amperes/m²), A is the constant of source material, T is temperature in kelvin (K), ϕ is work function, and k is the Boltzmann constant. The electron beam brightness is defined as emission current density per solid angle, or amperes/(m².steradian) in SI units. It is a rule of thumb that the LaB₆ emission has a brightness of more than 10 times that of the W filament source, and field emission is about 1,000 times brighter than that of LaB₆.

After electrons are emitted, the beam is accelerated by an electric field. Electrons travel down the column at a speed governed by a relativistic kinetic energy equation, $E = \frac{1}{2}mv^2$, where E ($= eV$) is energy in electron-volts that each electron carries, m is electron relativistic mass, and v is electron velocity.

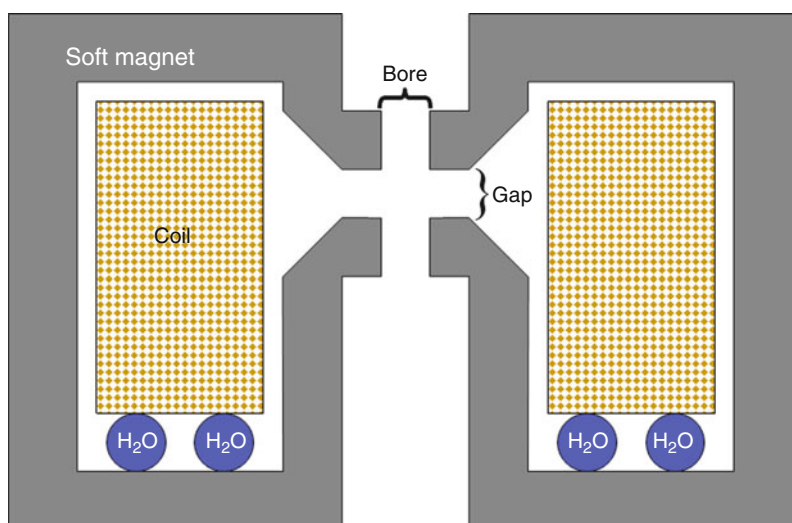
Condenser Lens

In electron optics, electromagnetic lens is employed for electron trajectories. Figure 2 is a schematic diagram of electromagnetic lens consisting of a tightly wound coil of Cu wire contained within a cylindrical soft-iron shroud. A chilled water cooling loop is usually necessary for heat dissipation. When electron beam passes through the bore,

a strong magnetic field exerts Lorentz force on electrons, causing them to travel in a spiral path. Rigorous treatment of electron trajectory equations for an electromagnetic lens concludes that for a given lens setting, one-to-one correlation holds true between an object and its image. SEM employs an electromagnetic lens to function as a condenser lens for converging or focusing an electron beam.

The electromagnetic lens has various types of aberrations, just as the conventional lens does. These aberrations include spherical aberration, astigmatism, chromatic aberration, coma, and distortion, among which only astigmatism can be rationally corrected by an operator. All other aberrations are associated with lens design and intrinsic lens physics. Individual aberration coefficients cannot be changed by an operator, even though the extent of aberration effects can be controlled by selecting optimal operation parameters. Unlike conventional light optics, where the lens can be concave or convex, an electromagnetic lens is always convex. Another important feature is that its focal length can be adjusted by changing lens current.

There are typically two to three condenser lenses in an SEM column. One is for focusing an electron beam on the specimen surface; the other one or two lenses in the up-column are used for adjusting electron probe current and spot size. A modern FEG SEM may employ a design of single condenser in the up-column, primarily due to the fact that the FEG itself provides a very small virtual source size at the first beam crossover underneath the gun tip.



Scanning Electron Microscopy (SEM), Fig. 2 Schematic diagram of electromagnetic lens

Scan Coils

Scan coils typically locate in condenser lens for focusing and deflect the electron beam in x and y directions at desired scan speed and raster size. Image magnification is inversely proportional to scan length. In an SEM, the adjustment of beam scanning is independent of major lens settings.

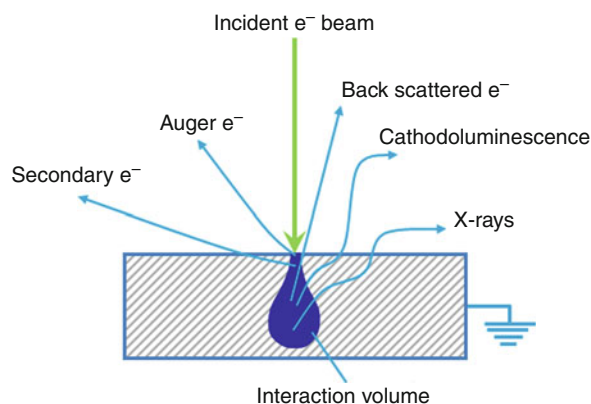
Detectors

There are three general types of electron detectors for SEM: Everhart-Thornley detector, Robinson (annular) detector, and solid state diode detector. These detectors are either mounted around the specimen chamber or in the beam path to collect and transfer signals for electronic or optical display and recording (see details in next section).

Imaging Mechanism

Electron Beam and Specimen Interactions

After electrons are emitted from electron gun, they are accelerated to an energy level typically in the range of 0.1–35 keV for modern SEM. The focused electron beam then strikes the specimen. Interactions between electron beam and specimen result in various secondary signals as schematically indicated in Fig. 3, including secondary electrons, backscattered electrons, characteristic X-ray photons, and cathodoluminescence. The rest of this section will briefly discuss a few commonly used signals for SEM imaging and analysis.



Scanning Electron Microscopy (SEM), Fig. 3 Interactions between electron beam and specimen

Secondary electrons: Secondary electrons result from the ionization of specimen atoms caused by incident electrons. The kinetic energy of secondary electrons is usually less than 50 eV. One or more Everhart-Thornley detectors are used in SEM to collect this type of signal to obtain the topographical image of the specimen (see details in section “[Electron image contrast](#)” below).

Backscattered electrons: When incident electrons collide head-on with atoms in the specimen, they are scattered “backwards” at approximately 180° .

These backscattered electrons retain all or most of the energy that incoming electrons carry. The yield of backscattered electrons is proportional to atomic number of the element. Therefore, imaging utilizing backscattered electrons reveals the elemental distribution at the surface. Backscattered electrons are collected for imaging by a Robinson detector or a solid state diode detector. In addition to the atomic number, backscattered electron yield and scattering angle are also sensitive to the crystal structure, and a technique called electron backscattered diffraction (EBSD) can offer crystal analysis.

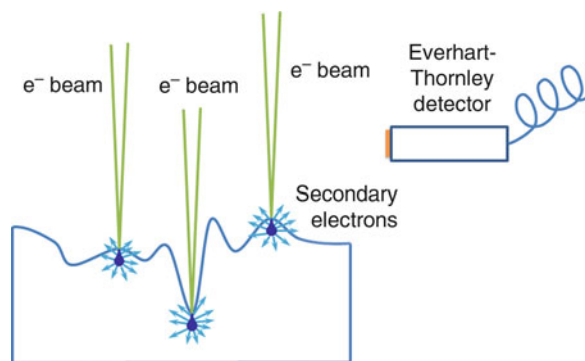
X-rays: A secondary electron from ionization leaves a vacancy in the inner shell of the electron orbit. A higher shell electron from the same atom can then “fall” to fill the vacancy. This creates an energy surplus in the atom, frequently resulting in the emission of an X-ray photon. These X-ray photons have characteristic energy levels specific to the elements in the specimen, which can be used for compositional analysis.

Image Formation and Spectrum

As an electron beam scans across a specimen and generates various secondary signals, one or more detectors can be set to collect specific signals or a combination of signals. The sensor or device in the detector transfers these signals for amplification and processing before being displayed and recorded. The image display is usually achieved by synchronizing and superimposing the scanned location coordinates and the processed signal intensity. Depending on the nature of dominant signals, images so formed can usually be called secondary electron (SE) image or backscattered electron (BSE) image.

Image magnification is easily set by scan unit. As the image display on monitor usually has a fixed size, the magnification is inversely proportional to the scan length on a specimen. It is also important to notice that the adjustment to the scan unit is independent of the focusing unit (last condenser lens), which means that if an image is in focus at a higher magnification, reducing the magnification on the same area does not affect the image focus. This feature offers a practical convenience for operating an SEM.

For an X-ray spectrum, X-ray photon counts are plotted against the respective photon energy in keV. For X-ray mapping, counts of the X-ray photon with specific energy can be superimposed with scanned location coordinates and/or with images. The associate technique is called X-ray energy dispersive spectroscopy (XEDS).



Scanning Electron Microscopy (SEM), Fig. 4 Secondary electron signal intensity variations due to surface topography

Electron Image Contrast

Image contrast in SEM fundamentally originates from signal intensity variation among adjacent locations. SEM image contrast is achieved in two major mechanisms: secondary electron signal intensity variation due to surface topography and backscattered electron signal intensity variation due to elemental distribution. The dependence of secondary electron yield on surface topography is illustrated in Fig. 4. As shown, secondary electrons formed at “valleys” may not be able to survive the path to get into the detector, while secondary electrons generated at “peaks” can be easily collected. It is worth noting that secondary electrons usually have low kinetic energies (<50 eV) and the Everhart-Thornley detector can normally be set with a positive bias to help collect the signal. Secondary electron images, therefore, reflect the topographical change. The contrast from backscattered electrons, on the other hand, indicates elemental distribution in a specimen.

Major Operation Parameters Impacting Image Resolution and Signal-To-Noise (S/N) Ratio

SEM instrument resolution is defined as the smallest distance between two discernible points under optimized operation conditions. Modern high-end SEM reaches sub-nanometer resolution, and usually more than one resolution value is specified for different operation conditions. SEM image resolution is ultimately a function of the smallest achievable beam diameter when it is focused on a specimen and the associated interaction volume, both of which are related to some major operation parameters including accelerating voltage, working distance, probe

current or spot size, and aperture size. These parameters at the same time affect the signal-to-noise (S/N) ratio.

Accelerating voltage: For metals and alloys, higher voltage usually improves image resolution due to two major factors: (a) less lens aberration effects to the higher energy beam, and (b) better probe current adjustment for smaller spot size and adequate S/N ratio. However, for semiconductor and polymer samples, higher energy beam frequently corresponds to larger interaction volume and higher probability of charging/discharging, in which case selecting a smaller accelerating voltage from less than 1–5 kV is usually beneficial to the image resolution.

Working distance: The working distance (WD) is the distance in millimeters between the last lens and the specimen surface. Decreasing WD always reduces the spot size and improves the image resolution. However, due to the detection solid angle limitations, there are optimal WDs associated with some detectors. Also, smaller WD corresponds to smaller depth of field. Here, the depth of field refers to the range of surface topographical height variation in a raster within which an overall image focus is achievable under given magnification and operation conditions.

Probe current: For SEM, probe current and spot size are practically viewed as equivalent and interchangeable

nomenclatures. Smaller probe current or spot size increases the achievable image resolution, but reduces S/N ratio at the same time. Smaller spot size also increases the depth of field.

Aperture size: The effects of aperture size are similar to those of the probe current.

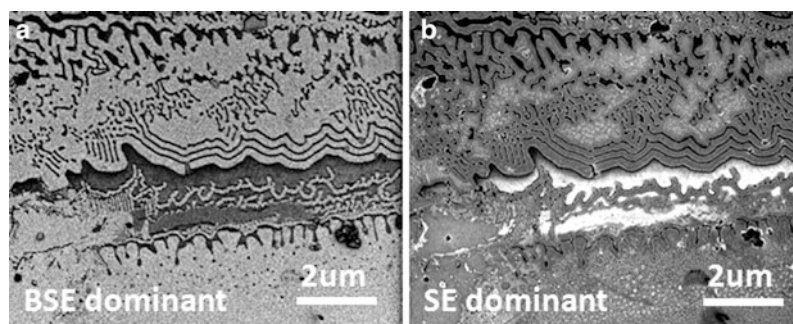
Key Applications

Imaging Using Secondary Electrons

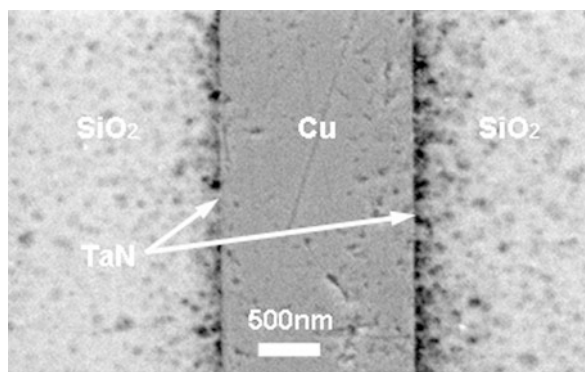
As discussed above, depending on samples and purposes of analyses, SEM operation parameters need to be optimized to achieve best results. For magnifications at $\times 100,000$ or above, a WD of 3 mm or smaller is usually necessary to obtain desired spot size and image resolution. Due to detector solid angle variation associated with decreased WD, in-lens detectors are usually more suitable for acquiring short WD and high-magnification images. **Figure 5** is an image acquired using JSM-7400F operating at 0.8 kV and WD 1.9 mm and utilizing an in-lens detector for secondary electron collection. Pore channels of ~ 10 nm in diameter are noticeably revealed. As the material contains a single uniform phase, the contrast is from the topographical variation.



Scanning Electron Microscopy (SEM), Fig. 5 Secondary electron image of a zeolite showing pore channels of about 10 nm in diameter



Scanning Electron Microscopy (SEM), Fig. 6 Images acquired on a layered perovskite $\text{La}_{2-x}\text{Sr}_x\text{CoO}_4$ using (a) predominant BSE and (b) predominant SE



Scanning Electron Microscopy (SEM), Fig. 7 SEM image of corrosion defects at $\text{SiO}_2/\text{TaN}/\text{Cu}$ -interconnect

Imaging Using Backscattered Electrons

Imaging using predominant backscattered electrons shows the contrast due to elemental distributions. Figure 6 is an example, where (a) is acquired using predominant BSE, and for comparison purpose, (b) is an image using SE. The image contrast in (a) is primarily due to compositional variations, and larger atomic number elements (La, Sr, and Co in this case) concentrate in the brighter areas. The contrast in (b) is mainly due to surface topographic variations. The brighter areas of resin inclusion located in the middle of the image are, however, the result of surface discharging.

Defect Review of Semiconductor Wafers after Chemical Mechanical Polishing (CMP)

In modern semiconductor chip fabrication, chemical mechanical polishing (CMP) has become an indispensable process. CMP exploits simultaneous and synergistic

chemical and tribological interactions between wafer, polishing slurry, and pad to achieve global wafer planarization. For product development or quality control, post-CMP wafer defect review by SEM is usually necessary. Figure 7 shows an image of polishing defects at the interfaces between SiO_2 , TaN, and Cu-interconnect, which include scratches, pits, and material losses of SiO_2 and TaN from corrosion and grinding.

X-Ray Energy Dispersive Spectroscopy (XEDS)

X-ray energy dispersive spectroscopy (XEDS) is a common attachment to SEM for compositional analysis based on the characteristic X-ray emission. In recent years, an increased interest in this technique is due to the development and perfection of silicon drifted detector (SDD) technology, which promises fast and accurate analyses. XEDS allows spectrum acquisition, elemental line scan profiling, and elemental mapping.

Cross-References

- [Atomic Force Microscopy \(AFM\)](#)
- [Auger Electron Spectroscopy \(AES\)](#)
- [Electron Energy Loss Spectroscopy \(EELS\)](#)
- [Scanning Tunneling Microscope \(STM\)](#)
- [Secondary Ion Mass Spectroscopy \(SIMS\)](#)
- [Transmission Electron Microscopy \(TEM\)](#)
- [X-ray Photoelectron Spectroscopy \(XPS\)](#)

References

- J. Goldstein, D. Newbury, D. Joy, C. Lyman, P. Echlin, E. Lifshin, L. Sawyer, J. Michael, *Scanning Electron Microscopy and X-Ray Microanalysis*, 3rd edn. (Springer, New York, 2003)
- D.B. Williams, C.B. Carter, *Transmission Electron Microscopy: A Textbook for Materials Science*, 2nd edn. (Springer, New York, 2009)

Scanning Force and Friction Microscopy (SFFM)

► Friction Force Microscopy (FFM)

Scanning Transmission Electron Microscopy (STEM)

► Characterization of Microstructures by Transmission Electron Microscopy

Scanning Tunneling Microscope (STM)

YI HE¹, SHENG FU CHEN^{1,2}, QIUMING YU¹

¹University of Washington, Seattle, WA, USA

²Zhejiang University, Hangzhou, People's Republic of China

Definition

STM is a scanning probe technique capable of exploring surfaces at the atomic scale and resolving the position of single atoms or molecules based on the concept of quantum tunneling of electrons between an atomically sharp tip and a conducting solid surface.

Scientific Fundamentals

Tunneling Current

The tunneling of electrons from one conductor to another through an insulator is a quantum mechanical phenomenon that has been known for more than 80 years. Although classically the electrons involved do not have enough energy to surmount the potential energy barrier of the insulator, quantum mechanically they have a small probability of reaching the other side of the barrier if the height and width of the barrier are sufficiently small. If two conductors are placed within a few angstroms of each other, the exponentially decaying electronic wave functions of the two conductors can overlap sufficiently for the electrons to have a probability of “tunneling” from one conductor to the other. If a potential is applied between

the two conductors, the resulting tunneling current can be measured. For two flat, parallel electrodes, this tunneling current I_T has an exponential dependence on the distance d between the two conductors (Fowler and Nordheim 1928):

$$I_T \propto \left(\frac{V_T}{d}\right) \exp(-A\Phi^{\frac{1}{2}}d)$$

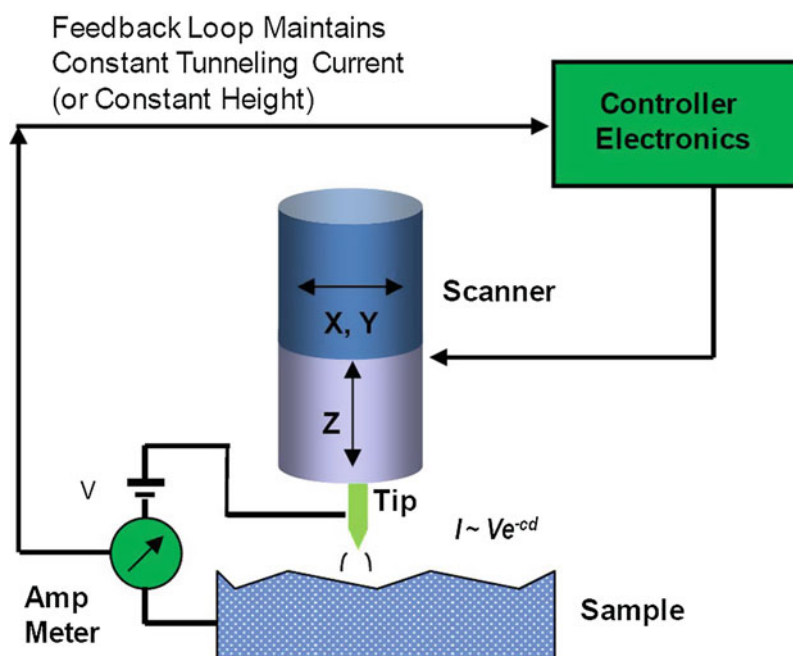
Here $A \approx 1.025 \text{ (eV)}^{-0.5} \text{ \AA}^{-1}$ for a vacuum gap, Φ is the average of the two electrode work functions, d is the distance between the electrodes, and V_T is the applied voltage. With work functions of a few eV, I_T changes by an order of magnitude for every angstrom variation of d . Therefore, the current may flow between electrodes when the distance is sufficiently small.

According to Tersoff and Hamann's theory (Tersoff and Hamann 1983, 1985), the tunneling current between a metallic tip and a metallic surface is approximately proportional to the surface local density of states (LDOS) at the Fermi level, E_F , evaluated at the location of the tip. While this simple approximation is subject to certain assumptions, such as small bias, low temperature, and a spherical tip represented solely by an s-wave function, the interpretation of STM images in terms of the surface LDOS at (or near) E_F has met with overwhelming success.

For surfaces decorated with relatively weakly interacting molecules, in contrast to the assumptions underlying Tersoff and Hamann's theory, larger bias voltages (typically on the order of 1 V) and higher temperatures (usually approximately room temperature) are employed for imaging such systems, thus expanding somewhat the energy range of surface states involved in tunneling. In addition, when considering the density of states of the overall system, the molecular orbital associated with the adsorbate must be taken into account.

Basic Setup

Similar to other types of scanning probe microscopes (SPMs), an STM contains a sharp probe that scans relative to a sample surface using precisely controlled voltages applied to piezoelectric elements, while the motion of the tip (or sample) normal to the sample surface is controlled by a feedback loop (see Fig. 1). What sets an STM apart from other SPMs is the use of the tunneling current between the tip and the sample as the signal for the feedback loop. Due to this required tunneling current, STM studies are limited to conductive samples. The tunneling current has an exponential dependence on



Scanning Tunneling Microscope (STM), Fig. 1 Schematic of the scanning tunneling microscope. The tip is shown mounted on a piezoelectric tripod with three orthogonal scanners marked x, y, and z. A bias is applied between the tip and the substrate. The feedback loop maintains either constant-current or constant-height operation mode

the distance between the tip and the sample. This enables a spatial resolution rarely achieved in other SPMs. The bias voltage applied between the tip and the sample typically varies from millivolts for bare metal or graphite surfaces to 1–2 V for semiconductor substrates and molecular adsorbates. Tunneling currents are usually in the pico-ampere to nano-ampere range. As no force measurement is taking place, the STM probe is not attached to a cantilever. Instead, the probe is the atomically sharp tip of a metal wire. Typically, either mechanically formed Pt/Ir tips or electrochemically etched Pt/Ir or W tips are employed.

Operation Modes

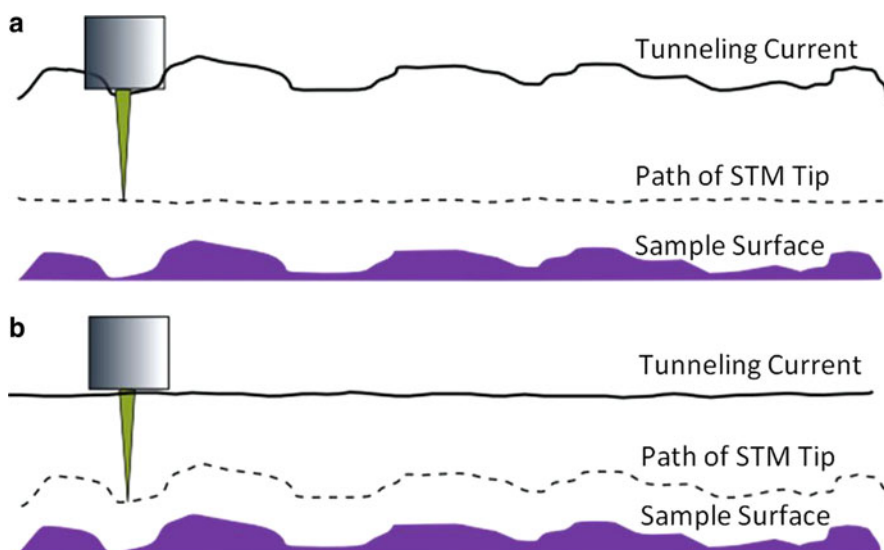
There are two major operation modes for STM: the constant-current mode and the constant-height mode (see Fig. 2). In the constant-current mode, a feedback circuit measures the tunneling current I_T and adjust the z position of the tip continuously in order to keep the current constant and equal to the reference value, I_{ref} , at the chosen bias voltage. Constant-current images are often very close to the topography of a surface whose variation of the local density of states at the Fermi level is small. In

the constant-height mode, the feedback circuit is turned off and the tip is scanned over the surface at nearly constant z position while the tunneling current is monitored at the chosen bias voltage.

Each mode has advantages and disadvantages. Constant-height mode is faster because the system does not have to move the scanner up and down, but it provides useful information only for relatively smooth surfaces. Another advantage of the constant-height mode is that the high scanning rates that can be achieved make it possible to examine some dynamic processes. Constant-current mode can measure irregular surfaces with high precision, but the measurement takes more time.

Key Applications

STM can provide atomic resolution images of metal and semiconductor surfaces as well as the adsorption of molecules. It can be operated in ultrahigh vacuum (UHV), ambient, or liquid conditions. Early STM work focused mainly on the clean, bare surfaces that exist under ultrahigh vacuum (UHV) conditions to study the structural and electronic properties of metal and semiconductor surfaces and provide ideal systems for formulating and



Scanning Tunneling Microscope (STM), Fig. 2 Schematic of the operation modes for STM: (a) constant-height mode, (b) constant-current mode

testing theories of tip-sample interactions and electron transport. UHV STM was also applied to study the kinetics and molecular interactions of adsorbates on metal and semiconductor surfaces at the atomic level. After the first commercial STM operating in ambient and low-current conditions was introduced in 1986, STM research has greatly expanded. The ability to perform STM work in air, under solutions, or within electrochemical cells has enabled the applications of STM to study self-assembled monolayers, biomolecules on conducting surfaces, and chemical reactions on surfaces. The high resolution that can be achieved with relative ease in STM images sets this technique apart from other SPMs and has allowed numerous STM studies to address the structure and dynamics of self-assembled monolayers in exquisite detail, and to study electron transport with submolecular resolution. Although STM is a unique and powerful technique for resolving atomic-scale surface structures, it does not provide direct information on chemical identities of the atoms constituting the surface structure.

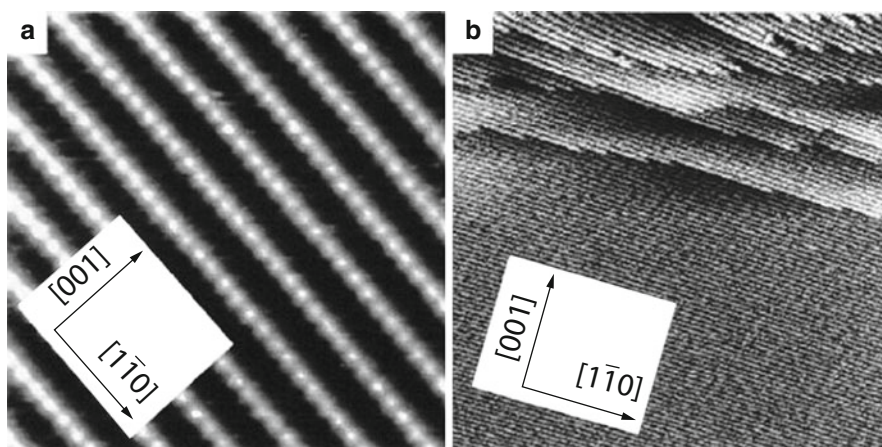
Reconstruction of Metal Surfaces

The creation of a surface means that the coordination of the atoms and the electronic structure in the surface region is changed relative to the bulk. For a few selected metals such as Au, Pt and Ir, the atomic positions are drastically rearranged in the surface region, from which new two-dimensional (2D) surface structures with significantly altered atomic densities originate. When this

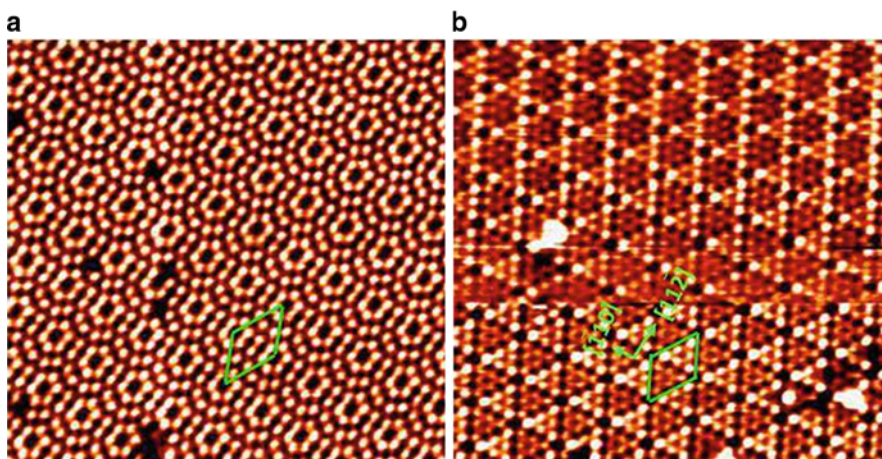
occurs, the surfaces are said to be reconstructed. STM has given unprecedented new insight into the reconstructions of these clean metal surfaces. For example, for the open (110) surface of gold, the reconstructions originate from the formation of a (1×2) missing row structure in which every second $[1\bar{1}0]$ close-packed surface atom row is missing (see Fig. 3), thereby forming (111) microfacets, resulting in surface structures with a lower surface energy than the unreconstructed (1×1) surface structures. Furthermore, the dynamic process of the surface reconstruction can be studied from time-lapsed STM images.

Reconstruction of Semiconductor Surfaces

The Si(111) 7×7 structure is a special example of a semiconductor surface for which STM images directly provide geometric information about the positions of surface atoms. STM images of the Si(111) 7×7 surface reveal 12 “topographic” maxima per unit cell (see Fig. 4). These topographic maxima can be attributed to the dangling bonds on the adatoms of the (7×7) surface structure. There are 12 adatoms per (7×7) unit cell. Each adatom ties up three dangling bonds from the underlying atomic layer, leading to a single dangling bond on each adatom due to the fourfold coordination of silicon. The dangling bonds on the adatoms are partially filled and therefore contribute to both empty and filled states. The positions of the observed maxima do not depend on the polarity of the applied bias voltage (i.e., the maxima of the empty and filled states are spatially coincident.)



Scanning Tunneling Microscope (STM), Fig. 3 STM topographs of Au(110) showing the (1×2) missing row reconstruction: (a) $(84 \times 84 \text{ Å}^2)$, and (b) $(800 \times 800 \text{ Å}^2)$, in which case several atomic steps exist at the top of the image (Besenbacher 1996)



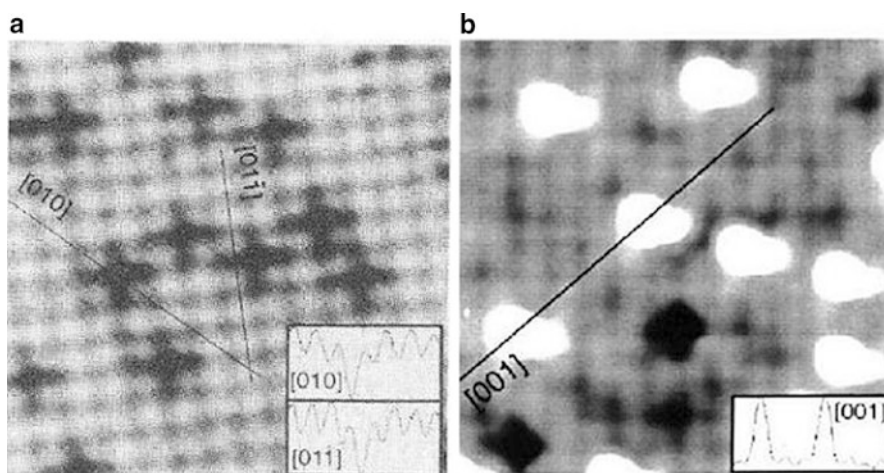
Scanning Tunneling Microscope (STM), Fig. 4 STM topographic images of clean Si(111) 7×7 surfaces. (a) Topograph $(300 \times 300 \text{ Å}^2)$ of the unoccupied states (empty states) obtained with the sample biased at +1.8 V and tunneling current of 0.2 nA. The unit cell is outlined and 12 adatoms are clearly visible. (b) Topograph $(300 \times 300 \text{ Å}^2)$ of the occupied states (filled states) obtained with the sample biased at -3.0 V and tunneling current of 0.1 nA. The stacking fault and the differences between corner and center adatoms are visible

Although STM images of the Si(111) 7×7 surface directly reveal the positions of the adatoms, as already discussed, there are still significant electronic contributions to the observed image contrast. For instance, STM images obtained with positive sample bias voltage usually reveal 12 adatoms of equal height in each unit cell (Fig. 4a). In contrast, in STM images obtained with negative sample bias voltage, the adatoms in the faulted half of the unit cell appear “higher” than those in the

unfaulted half (Fig. 4b). Furthermore, the adatoms located next to a corner hole (“corner adatoms”) appear slightly “higher” than the central adatoms. By increasing the sample bias, the rest atoms will become visible as well, in addition to the adatoms.

Adsorbates on Metal Surfaces

The image contrast can be used to resolve adsorbates on surfaces. From a simple perturbative approach, the



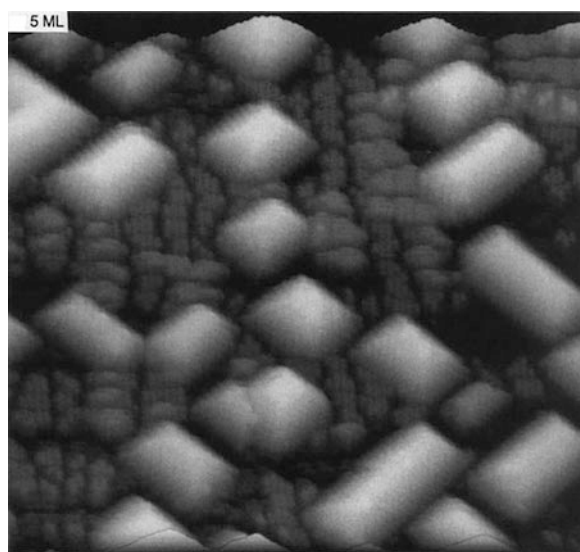
Scanning Tunneling Microscope (STM), Fig. 5 STM images for (a) carbon atoms chemisorbed on Ni(100) ($32 \times 42 \text{ \AA}^2$) (Klink et al. 1993), and (b) sulfur atoms chemisorbed on Ni(100) ($35 \times 38 \text{ \AA}^2$). (Besenbacher 1996) The lower insets in (a) and (b) show the corrugation along (a) $\langle 010 \rangle$ (z scale 0.36 \AA) and $\langle 011 \rangle$ (z scale 0.23 \AA), (b) $\langle 001 \rangle$ (z scale 0.4 \AA), respectively

contrast of single, individual adsorbates in an STM image depends on the change induced by the adsorbate in the LDOS near E_F . An adsorbate appears as a protrusion if it causes an increase in the LDOS at E_F ; conversely, it is imaged as a hole if the LDOS is depleted relative to the clean surface.

Figure 5 shows STM topographic images of low concentration carbon and sulfur adsorbates on Ni(100). The carbon atoms appear as depressions with a depth of $\sim 0.3 \text{ \AA}$ (Fig. 5a) while sulfur is imaged as protrusions with a height of $\sim 0.4 \text{ \AA}$ (Fig. 5b) relative to the Ni(100) 1×1 surface. This indicates that the adsorption of carbon and sulfur atoms on Ni(100) causes a depletion and increase in the LDOS at E_F , respectively.

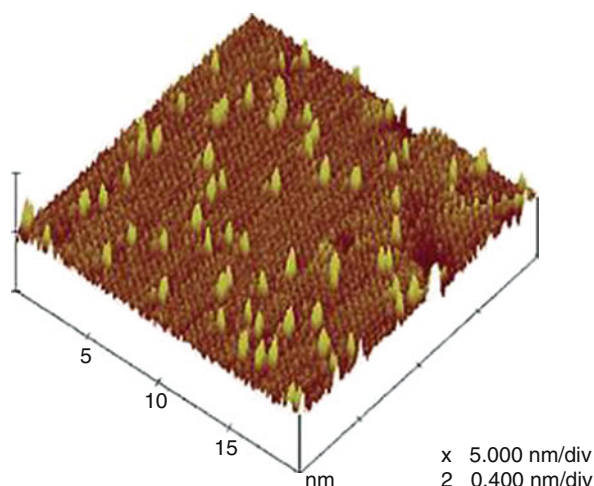
Epitaxial Growth on Semiconductor Surfaces

Another important application of STM is to determine the growth mode and study the structural properties of the epitaxial growth of semiconductors by molecular beam epitaxy (MBE) or chemical vapor deposition (CVD). For example, heteroepitaxial growth of Ge on Si(100) surfaces has been investigated by STM. In the submonolayer and low coverage range (less than three monolayer (ML)), the growth is layer-by-layer and the Ge dimers can be resolved in STM images (Fig. 6). In the intermediate coverage range (more than 3 ML), STM images show three-dimensional small islands with (111) facets (“hut” clusters) (Fig. 6),



Scanning Tunneling Microscope (STM), Fig. 6 STM image of 5 ML Ge deposited onto Si(100) at 650K. (Knall and Pethica 1992) Both the characteristic “hut” clusters and Ge dimers are shown. The size of the image is approximately $40 \times 40 \text{ nm}^2$. The image was obtained with the sample biased at -2.0 to -3.0 V and tunneling current of $1\text{--}3 \text{ nA}$

which are characteristic of a Stranski–Krastanov growth mode, indicating lattice mismatch between silicon and germanium crystals.



Scanning Tunneling Microscope (STM), Fig. 7 Low-current STM topographic images of mixed self-assembled monolayers on Au(111) formed from a solution of 11-mercaptoundecanoic acid and decanethiol (1:9) at 65°C with a bias voltage of 1.02 V and a set point current of 1.10 pA (Li et al. 2003)

Self-Assembled Monolayers

Thiols chemisorbed onto Au(111) are one of the self-assembled systems most extensively investigated using STM. The chemisorption of thiols on gold is an irreversible process driven by the formation of the strong Au-S bond. Intermolecular interactions and the structure of the Au(111) surface also play a role during the self-assembly. Dense monolayers are formed, in which only the thiol head-group is in contact with the gold surface. In the case of alkane thiols, a commensurate ($\sqrt{3} \times \sqrt{3}$)R30° overlayer structure is formed, with additional superstructures depending on chain length and functionalization. Figure 7 shows an example of an ambient high-resolution STM image, wherein the ordered structure of a mixed thiol layer can be seen clearly. Both molecular species of interest are incorporated in the monolayer and can be distinguished in the STM image.

References

- F. Besenbacher, Scanning tunnelling microscopy studies of metal surfaces. Rep. Prog. Phys. **59**(12), 1737–1802 (1996)
- R.H. Fowler, L. Nordheim, Electron emission in intense electric fields. Proc. R. Soc. London Ser. A-Contain. Pap. Math. Phys. Character **119**(781), 173–181 (1928)
- C. Klink, L. Olesen et al., Interaction of C with Ni(100) – atom-resolved studies of the clock reconstruction. Phys. Rev. Lett. **71**(26), 4350–4353 (1993)
- J. Knall, J.B. Pethica, Growth of Ge on Si(100) and Si(113) studied by STM. Surf. Sci. **265**(1–3), 156–167 (1992)

- L.Y. Li, S.F. Chen et al., Molecular-scale mixed alkanethiol monolayers of different terminal groups on Au(111) by low-current scanning tunneling microscopy. Langmuir **19**(8), 3266–3271 (2003)
- J. Tersoff, D.R. Hamann, Theory and application for the scanning tunneling microscope. Phys. Rev. Lett. **50**(25), 1998–2001 (1983)
- J. Tersoff, D.R. Hamann, Theory of the scanning tunneling microscope. Phys. Rev. B **31**(2), 805–813 (1985)

Schallamach Waves

- [Contacts Involving Wave Propagation](#)

Science of Rubbing

- [Tribology](#)

Science of Surfaces in Contact and Relative Motion

- [Tribology](#)

Screening Wear Test

- [Polymers in Biotribology](#)

SE – Surface Energy

- [Interfacial Energy](#)

Seal

- [Mechanical Seals](#)

Secondary Ion Mass Spectroscopy (SIMS)

SUKBAE JOO, HONG LIANG

Department of Mechanical Engineering, Texas A&M University, College Station, TX, USA

Definition

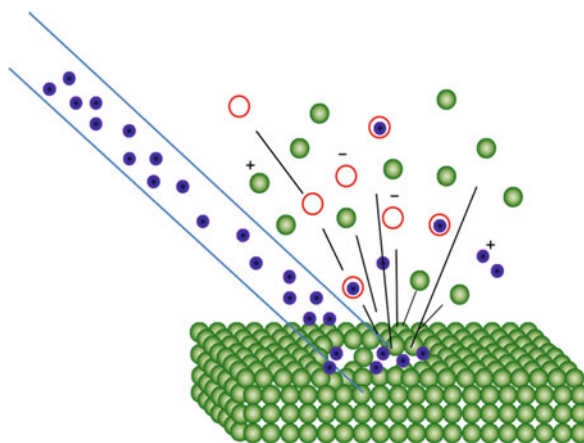
SIMS, secondary ion mass spectroscopy, is a surface chemical analysis technique for solid materials. As its name indicates, a specimen is bombarded with a primary ion beam and the secondary ions are collected using a detector – a spectrometer. The secondary ions provide information on the elemental, molecular, and isotopic composition of a material's surface. SIMS is one of the most sensitive techniques for surface analysis.

Scientific Fundamentals

In 1910, Joseph John Thomson, a British physicist, observed the emission of secondary ions followed by the bombardment of a metal surface (Thomson 1910). It took almost 30 years after his observation to enable the qualitative and quantitative analysis of secondary ions. In 1949, Herzog and Vieböck built the first prototype of SIMS and analyzed secondary ions from metals and oxides (Herzog and Vieböck 1949). It was in the 1960s that two types of practical SIMS were developed: one by Castaing and Slodzian in 1960 and another by Liebel and Herzog in 1967.

Principles

For SIMS, a solid sample surface is bombarded with an accelerated high energy (1–30 KeV) primary ions such as Cs^+ , O_2^+ , O^- , Ar^+ , N_2^+ , and Ga^+ . Subsequently, electrons, photons, and secondary particles are generated from the surface, as illustrated in Fig. 1. The secondary particles can be positively, negatively, or neutrally charged species, which have kinetic energy in a range from zero to several hundred electron voltage. This process is called *sputtering*. When the primary ions interact with sample surface, several monolayers up to approximately 10 nm can be collided by collision cascade. The collision cascade in solids is generated when an atom is displaced from its lattice by a high energy source and moves through the materials, producing another displacement continuously (Parkin 1990). The emission of all secondary particles is originated from this collision cascade phenomenon.



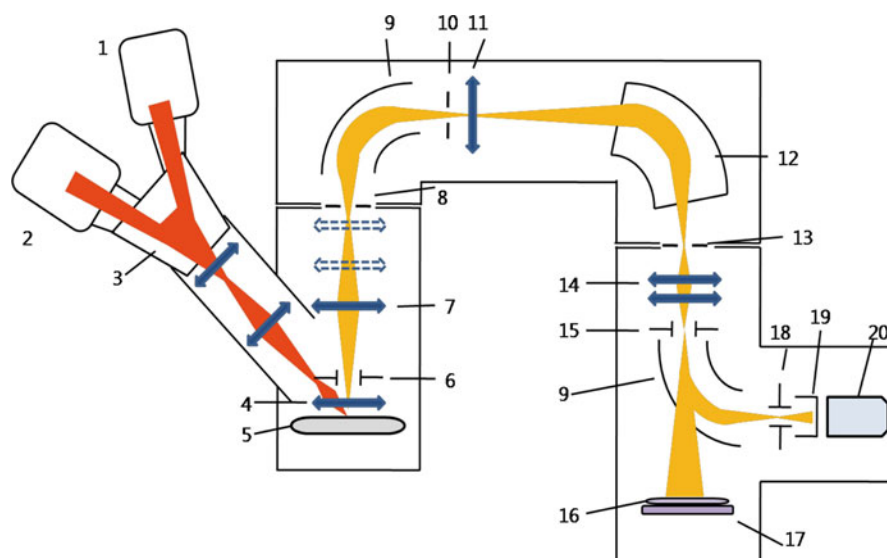
Secondary Ion Mass Spectroscopy (SIMS), Fig. 1 Schematic of ion beam sputtering

Due to the nature of sputtering, the removal of sample material is not negligible, even though the amount is small; it sometimes causes sputter crater on the monolayer. This can be explained by monolayer lifetime and its formula given by (Vickerman et al. 1989):

$$t_m = \frac{N_s}{6.2 \times 10^{18} I_p Y} \quad (1)$$

where N_s is the density of atoms in the surface and its unit is cm^{-2} ; Y is the ratio of the number of secondary particles to the number of ions primarily sputtered on the target, which usually lies between 0.1 and 10 (Anderson and Bay 1981); and $6.2 \times 10^{18} I_p$ is the number of primary ions directed to the surface and its unit is $\text{cm}^{-2} \cdot \text{s}^{-1} \cdot \text{A}^{-1}$ for the primary ion current I_p . Clearly, the sputter rates depend on primary beam intensity and material characteristics such as crystal orientation and topography.

While sputter yield provides information about the atomic characteristics of the material under sputtering, ionization probability offers more electrical and chemical information about the secondary ions. The ionization probability is also called ion yield, which indicates the fraction of secondary particles that become ionized. The ion yield varies over many orders of magnitude for the elements; the positive ion yield has a trend inversely dependent on ionization potential, and the negative ion yield has a similar trend, depending on electron affinity (Storms et al. 1977). Environmental factors and target materials also affect the ion yield. For example, with an increase of oxygen pressure in the analysis chamber, Si^+ ion yield increases by 3 orders of magnitude



Secondary Ion Mass Spectroscopy (SIMS), Fig. 2 The schematic of SIMS: 1. Cesium ion source, 2. Duoplasmatron ion source, 3. Primary beam mass filter, 4. Immersion lens, 5. Specimen, 6. Dynamic transfer system, 7. Transfer optical system, 8. Entrance slit, 9. Electrostatic sector, 10. Energy slit, 11. Spectrometer lens, 12. Electromagnet, 13. Exit slit, 14. Projection lenses, 15. Deflector, 16. Channel-plate, 17. Fluorescent screen, 18. Deflector, 19. Faraday cup, and 20. Electron multiplier

during Ar^+ bombardment of silicon (Maul and Wittmaack 1975). Another example is that the ion yield from oxidized metal surface is much higher than that from non-oxidized metal surface (Benninghoven 1975).

Electrochemical characteristics of primary ion can also be a factor that affects the ionization probability. It is well known that electronegative elements such as oxygen lead to dramatically higher positive secondary ion yield than an inert element such as argon does (Andersen 1970). The formation of strongly bonded compounds such as metal-oxygen bond decreases surface electron availability and its breakage by bombardment leads to positive charging of surface element. On the other hand, negative secondary ion yield increases with the bombardment of electropositive elements such as cesium. Implantation of cesium into the surface leads to decreased work function and, consequently, increased electron availability on the surface, which contributes to the generation of more negative ions (Andersen and Hinthorne 1972).

Regardless of the energy of the primary beam, secondary ions have kinetic energy distributions. As mentioned above, these distributions range from zero to several hundred electron voltage and generally peak

at low energies (1–10 eV). The energy distributions are significantly narrower for cluster ions than for elemental ions and shift to lower energy as cluster ion size increases (Lancaster et al. 1979). Analyzing the energy distribution against various clustered ions can lead to the identification of existing bonds through reactions if any, and thus chemical information.

Instrumentation

SIMS instrumentation can be roughly explained in two parts by process order. One is the secondary ion generation part and the other is the secondary ion measurement part. The schematic of SIMS instrumentation is shown in Fig. 2 (Valley et al. 1998).

The ion generation part consists of three components, including ion gun, primary ion column, and secondary ion extraction and transfer column. There are several types of ion guns according to the types of ion source. In SIMS, ion guns are used mainly with two ion sources; one is duoplasmatron and another is surface ionization source. In the duoplasmatron operation, a small amount of primary gas is let into the chamber and atoms react with electrons emitted from the cathode filament. When atoms interact with energetic electrons, ionization of atoms can

occur. As ionization proceeds, the number of ionization exceeds a critical value and forms plasma. This dense plasma is accelerated and compressed into an aperture in a planar anode and extracted as fairly high-speed ion beams form there. Inert gases such as Ar^+ or Xe^+ have been used as a primary gas for this operation; O_2 is currently the most frequently used primary beam source. In the surface ionization operation, when a low ionization potential element is on a high work function metal hot enough for the ion desorption from the surface, the element is desorbed as ions lead to the extractor (Wilson 1967). Cesium is a major primary ion source for the surface ionization.

After primary ions are extracted from the ion gun, this ion beam is transferred to the primary ion column. The column usually consists of a mass filter, electrostatic lenses, apertures, and deflectors. The mass filter eliminates impurities in the beam and electrostatic lenses and apertures control beam intensity and width. The controlled and modified beam is deflected to the desired area on the surface by deflector. The secondary ions by the bombardment of primary atoms are extracted and transferred in a high vacuum chamber. This high vacuum chamber has an extraction lens near sample and transfer lenses. The secondary ions are accelerated and transferred to the analyzing sector through the chamber.

The secondary ion measurement part consists of three components, an ion energy analyzer, mass analyzers, and detectors. The ion energy analyzer has electrostatic sector that controls ion beam by bending it in different degrees. As stated above, the secondary ion beam has energy distributions. The ion beams with different energy levels are bent in different degrees by the electrostatic sector. The high-energy ions are blocked by an energy slit. The low energy ions pass through the slit and are adjusted to the desired forms by the spectrometer lens.

The ion beam from the spectrometer lens is then transferred into mass spectrometer for mass analysis. There are normally three types of mass spectrometer according to different operating modes: quadrupole mass analyzers, magnetic sector mass analyzers, and time-of-flight (ToF) mass analyzers. According to these mass spectrometers, SIMS can be categorized into static and dynamic SIMS. Static SIMS and dynamic SIMS can be distinguished from each other depending on the sputtering rate. Static SIMS is operated by relatively low dose of the primary ions, thus only topmost atomic layer is sacrificed. On the other hand, dynamic SIMS is operated by relatively high dose of the primary ions consuming top

few monolayers. In early days, quadrupole has been used for static SIMS, but now it is used more for dynamic SIMS. Quadrupole generally consists of four rods of circular shape, which are connected together as two opposite pairs. Direct voltage is applied to one pair of rods and the opposite voltage is applied to the other pair. When ions go through the space among rods, most ions follow unstable oscillation and hit the rods, but ions with a single mass-to-charge ratio follow periodically stable oscillation and are transmitted to detectors. Magnetic sectors are most common for dynamic SIMS and its combination with electrostatic sector enables magnetic double focusing spectrometer. The magnetic sector plays a similar role as the electrostatic sector except for the different type of field. In a series of electrostatic and magnetic fields, the ions with various energy levels are separated and controlled when they pass the slit. Consequently, the high mass resolution can be obtained by this method. Time-of-flight (ToF) is the method measuring the flight time of ions until they arrive at detectors. The ions of different mass-to-charge ratio have different velocities. Therefore, the flight time of the ions varies. Since this method has high sensitivity compared with others, it is suitable for small area analysis and imaging.

SIMS detectors consist of four components, including an electron multiplier, a Faraday cup, and two ion image detectors. The electron multiplier is an ion counting detector and has continuous dynode (a series of electrode) structure in a vacuum tube. When a particle strikes one dynode, it induces the emission of secondary electrons. The secondary electrons then strike the next dynode and produce more secondary electrons. In this manner, a large number of electrons are generated on the metal anode. A Faraday cup is placed in front of the electron multiplier and protects the electron multiplier when the incoming ion signal is too high.

Two ion image detectors are microchannel plate arrays that are placed close to each other. The voltage across each channel plate can produce gains as high as 10^5 , and two combined plates can produce gains at least 10^6 . They are used to visualize the electron cascade and monitor the secondary ion beam.

Detection Limits

SIMS has a distinctly wider detection range than other similar techniques. The comparisons of detection limits, the range of detectable elements, depth resolution, and lateral resolution among SIMS, Auger electron spectroscopy (AES), and X-ray photoelectron

Secondary Ion Mass Spectroscopy (SIMS), Table 1 Detection limit comparison among similar techniques

	Typical application	Signal detected	Elements detected	Detection limits	Depth resolution	Lateral resolution
AES	Surface analysis, micro-area depth profiling	Auger electrons	Li~U	0.1~1 at.%	20~200 Å (profiling mode)	≥100 Å
					30 Å (surface analysis)	
XPS	Surface analysis of organic and inorganic materials, depth profiling	Photo electrons	Li~U	0.01~1 at.%	20~200 Å (profiling mode)	10 μm~2 mm
					10~100 Å (surface analysis)	
SIMS	Dopant and impurity depth profiling, surface and microanalysis, insulator analysis	Secondary ions	H~U	10 ¹² ~10 ¹⁶ at/cm ³ (ppb~ppm)	10~300 Å	≥30 μm (depth profiling)
						1 μm (imaging mode)

spectroscopy (XPS) are shown in Table 1. The SIMS detection limits for most elements are between 10¹² and 10¹⁶ atoms/cm³ and detectable elements range from hydrogen to uranium. Since SIMS uses sputtered ions, it offers relatively high resolution. Due to the sputtering on the sample, however, sample surface damage is inevitable. Static SIMS technique has been improved to reduce surface damage, but is still not perfect.

Key Applications

Element Identification

Since SIMS provides mass spectra based on mass-to-charge ratio and ion counts for certain elements, the identification of elemental and molecular ions is possible. Unlike other techniques for element identification, SIMS can detect the concentration of certain elements down to ppb. In semiconductor industries, as the device features shrink to less than 40 nm, contaminations on the silicon wafer need to be strictly controlled, which requires a very small detection limit. As mentioned above, SIMS is much more sensitive than AES and XPS, and elemental spectral overlap does not occur. Due to its sensitivity, SIMS is suitable for the analysis of extremely thin films such as oxide layers on metal. In particular, initiation of oxide layer can be analyzed by SIMS. Based on the element identifications, positional information of ions can be obtained and mapped. Image detectors indicate the position of arriving ions and their intensity and give

graphical information. An example of ion mapping is shown in Fig. 3. The images of Ga-implanted Si and In-implanted Si were mapped based on the different ion counts and these images clearly illustrate the difference in chemistry.

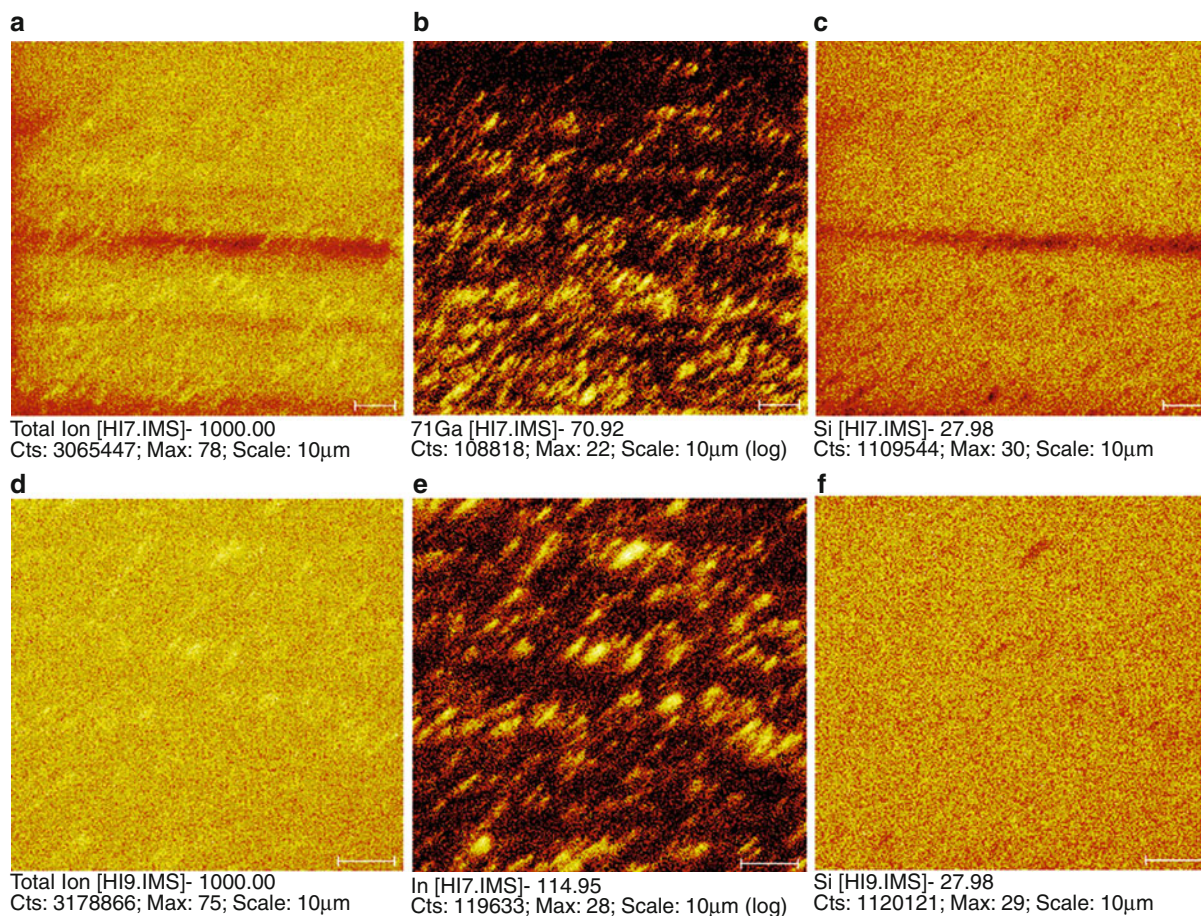
Defect Analysis

SIMS can also be used for defect analysis of semiconductor devices. Particularly, time-of-flight (ToF) SIMS is more suitable for defect and failure analysis than magnetic sector and quadrupole SIMS. The major strength of ToF-SIMS is the ability to detect and identify various inorganic and organic materials in one analysis. Furthermore, a full wafer (size up to 300 mm) can be analyzed at one time (Schnieders et al. 2010). Defect and failure can be also analyzed by ion imaging, and a visualized image has a lateral resolution around 1 μm.

Depth Profiling

SIMS depth profiling can be used to determine elemental concentrations of dopant and impurity atoms in a range from 10¹³ to 10²⁰ atoms/cm³ at depths of up to 10 μm (Vickerman et al. 1989). The advantages of SIMS depth profiling include high dopants/impurities sensitivity up to ppb, high depth resolution (1 nm), and quantitative analysis.

Dynamic SIMS is used for depth profiling. Depth profile can be obtained by measuring the secondary ion count of selected elements as a function of a time, or secondary count rate. The depth is determined



Secondary Ion Mass Spectroscopy (SIMS), Fig. 3 (a–c) ToF-SIMS chemical maps of the Ga on Si sample, and (d–f) ToF-SIMS chemical maps of the In on Si sample; (a) total ion map of Ga on Si, (b) Ga (m/z 71), (c) Si (m/z 28), (d) total ion map of In on Si, (e) In (m/z 115), and (f) Si (m/z 28). All maps originate from the same sample area (100 μm × 100 μm) (Courtesy of Dr. Kamal Soni from Corning)

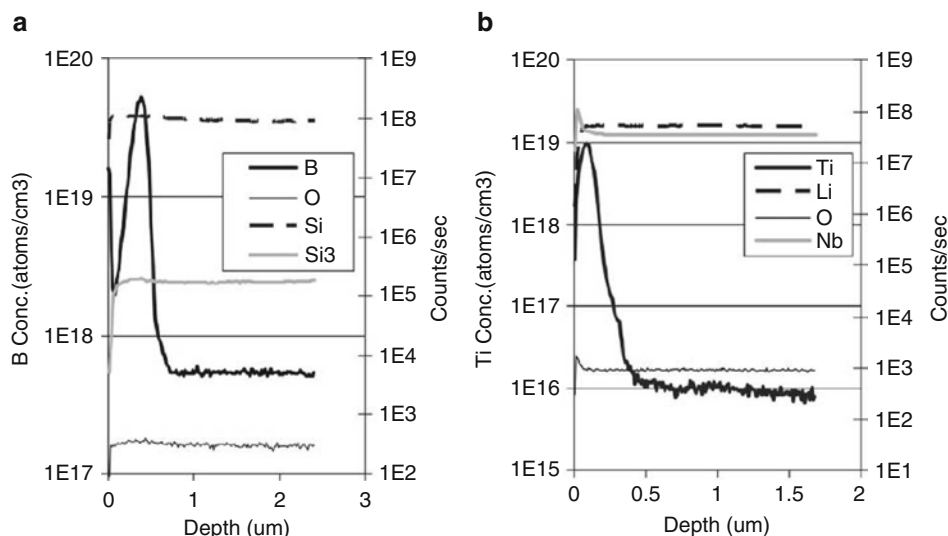
by measuring the crater depth with a surface profilometer. Secondary ion count rate can be translated to concentration of atoms using relative sensitivity factors (RSFs). Relative sensitivity factors are used for quantitative analysis by SIMS and explained by (Wilson et al. 1989):

$$C_E = \text{RSF} \cdot \frac{I_E}{I_M} \quad (2)$$

where C_E is elemental concentration, I_E is secondary ion intensity for element E, and I_M is secondary ion intensity

for matrix element. RSF are empirically obtained values and key factors for SIMS quantification. An example of depth profiling is shown in Fig. 4. Two depth profiles of ^{11}B implanted bulk SiO_2 and ^{48}Ti implanted LiNbO_3 provide the concentration of each dopant as well as other elements (Pivovarov et al. 2004).

If multilayer structures are complicated and each layer is thin, high depth resolution is essential to resolve abrupt interfaces and multilayer structures. The depth resolution can be improved empirically by changing the energy or the angle of incidence of the primary ion beam.



Secondary Ion Mass Spectroscopy (SIMS), Fig. 4 SIMS depth profile of (a) ^{11}B implanted in bulk SiO_2 and (b) SIMS depth profile of ^{48}Ti implanted in LiNbO_3 (Courtesy of Dr. Fred Stevie from Elsevier Ltd.)

Cross-References

- [Auger Electron Spectroscopy \(AES\)](#)
- [X-ray Photoelectron Spectroscopy \(XPS\)](#)

References

- C.A. Andersen, Analytic methods for the ion microprobe mass analyzer. Part II. Int. J. Mass Spectrom. Ion Phys. **3**, 413 (1970)
- C.A. Andersen, J.R. Hinthorne, Ion microprobe mass analyzer. Science **175**, 853 (1972)
- H.H. Anderson, H.L. Bay, Sputtering by particle bombardment I, in *Topics in Applied Physics*, ed. by R. Behrisch (Springer, Berlin, 1981), p. 145
- A. Benninghoven, Developments in secondary ion mass spectroscopy and applications to surface studies. Surf. Sci. **53**, 596 (1975)
- R.F.K. Herzog, F.P. Vieböck, Ion source for mass spectrography. Phys. Rev. **76**, 855–856 (1949)
- G.M. Lancaster, F. Honda, Y. Fukuda, J.W. Rabalais, Secondary ion mass spectrometry of molecular solids. Cluster formation during ion bombardment of frozen water, benzene, and cyclohexane. J. Am. Chem. Soc. **101**(8), 1951 (1979)
- J. Maul, K. Wittmaack, Secondary ion emission from silicon and silicon oxide. Surf. Sci. **47**, 358 (1975)
- D.M. Parkin, The displacement cascade in ceramic oxides. Nucl. Instrum. Methods Phys. Res., Sect. B **46**, 26 (1990)
- A.L. Pivovarov, F.A. Stevie, D.P. Griffis, Improved charge neutralization method for depth profiling of bulk insulators using O_2^+ primary beam on a magnetic sector SIMS instrument. Appl. Surf. Sci. **231–232**, 786 (2004)
- A. Schnieders, Full wafer defect analysis with time-of-flight secondary ion mass spectrometry, in *Advanced Semiconductor Manufacturing Conference (ASMC)*, (San Francisco, CA, 2010), p. 158
- H.A. Storms, K.F. Brown, J.D. Stein, Evaluation of a cesium positive ion source for secondary ion mass spectrometry. Anal. Chem. **49**(13), 2023 (1977)
- J.J. Thomson, Rays of positive electricity. Phil. Mag. **20**, 752–767 (1910)
- J.W. Valley et al., Ion microprobe analysis of oxygen, carbon, and hydrogen isotope ratios, in *Applications of Microanalytical Techniques to Understanding Mineralizing Processes*, ed. by M.A. Mckibben et al. Review in Economic Geology, vol. 7 (Society of Economic Geologists, Littleton, 1998), p. 73
- J.C. Vickerman, A. Brown, N.M. Reed, in *Secondary Ion Mass Spectrometry: Principles and Applications*, ed. by R. Breslow (Oxford University Press, Oxford, 1989), p. 10
- R.G. Wilson, Surface ionization ion sources. IEEE Trans. Nucl. Sci. **14**, 72 (1967)
- R.G. Wilson, F.A. Stevie, C.W. Magee, *Secondary Ion Mass Spectrometry: A Practical Handbook of Depth Profiling and Bulk Impurity Analysis* (Wiley, New York, 1989)

Seizure

- [Gear Contact Temperature and Scuffing Risk Analysis](#)

Selective Laser Surface Hardening

- [Laser Surface Hardening](#)

Self-Acting Bearings

- [Hydrodynamic Journal Bearing History](#)

Self-Acting Foil Gas Bearings

► Foil Gas Bearings

Self-Adjustment Bearings

► Self-Aligning Bearings

Self-Aligning Bearings

XIAOLAN AI

Timken Technology Center, The Timken Company,
Canton, OH, USA

Synonyms

Internal aligning bearings; Self-adjustment bearings; Self-alignment bearings; Spherical race bearings; Spherical roller bearings

Definition

Self-aligning bearings refer to a type of bearing designed and used to accommodate misalignment between the housing and shaft. Self-aligning bearings include self-aligning ball bearings, spherical roller bearings, toroidal roller bearings, and self-aligning thrust bearings

Scientific Fundamentals

Self-aligning bearings are designed for use in applications where misalignment or angular displacement between the shaft and bearing housing is anticipated. The misalignment can be accommodated either internally inside the bearing or externally outside of the bearing. [Figure 1](#) illustrates a spherical roller bearing where the misalignment is accommodated between rollers and raceways. The accommodation to the misalignment is achieved dynamically inside the bearing. [Figure 2](#) shows another example where the misalignment is accommodated externally at a spherical bearing seat. Although the term *self-aligning bearings* refers broadly to both alignment configurations, it is most commonly used for bearings with internal alignment capability.

Self-Aligning Ball Bearing

Self-aligning ball bearings are often constructed in a two-row arrangement, as illustrated in [Fig. 3](#). The bearing



Self-Aligning Bearings, Fig. 1 A spherical roller bearing having internal alignment capability (courtesy of the Timken Company)



Self-Aligning Bearings, Fig. 2 A cylindrical roller bearing having external alignment capability

contains an outer race ring having an outer raceway, an inner race ring having two grooved inner raceways, two rows of balls, and two separate cages, each retaining a row of balls. The outer raceway of this bearing is a portion of a sphere. Thus balls on the inner raceways are internally



Self-Aligning Bearings, Fig. 3 A self-aligning ball bearing in two-row arrangement

aligned with respect to the outer raceway to accommodate any misalignment, free from any moment loads. Unlike conventional ball bearings, the outer raceway in the self-aligning ball bearings does not conform well to the balls. This results in a reduced Hertzian contact area and load-carrying capacity (see ref (Hamrock 1981)). Self-aligning ball bearings have a relatively wide operating speed range and are less susceptible to lubrication conditions compared with other types of self-aligning bearings. Self-aligning ball bearings are designed to carry primarily radial loads and a limited amount of axial loads in both directions.

Application of Self-Aligning Ball Bearings

Typical applications are seen in mining machinery, power transmissions, heavy machinery, and textile machinery.

Spherical Roller Bearing

The most popular spherical roller bearings are in a two-row arrangement, as shown in Fig. 1. The bearing comprises an outer race ring, an inner ring having two raceways, and two rows of barrel-shaped rollers that are separated and retained by an integral cage. The outer race ring has a single outer raceway that is a portion of a sphere. This allows the inner ring and roller assembly to dynamically align inside the outer raceway, permitting a limited amount of angular displacement between bearing shaft and housing bore. Standard spherical bearings are designed to accommodate misalignment up to $\pm 1.5^\circ$. Rollers of a spherical roller bearing lie at an angle relative

to the axis of the bearing. A seating rib is presented between the two rows of rollers. The seating rib can be a part of or floating around the inner ring. The rollers are either symmetrical or asymmetrical barrel shaped and conform closely to both the inner and outer raceways. The high degree of conformity makes spherical roller bearings suitable for heavy-duty applications. Because of the non-zero contact angle, spherical roller bearings are capable of carrying a certain amount of thrust load in either direction along with the radial load.

Unlike cylindrical roller bearings or tapered roller bearings, true rolling motion at contacts between the rollers and raceways cannot be achieved with spherical roller bearings (see ref. (Harris 2001)). Thus, spherical roller bearings inherently have higher frictional torque than cylindrical roller bearings and are not suitable for high-speed applications.

Application of Spherical Roller Bearings

Spherical roller bearings are the most popular self-aligning bearings types. Applications are seen in a wide variety of industries and market segments. Typical applications include mining and construction equipment, rolling mills, paper mills, windmills, industrial gear drives, and material-handling equipment.

Toroidal Roller Bearing

Toroidal roller bearings were introduced to market by SKF, Inc. as non-locating bearings, also known as CARB bearings. The toroidal roller bearing, as shown in Fig. 4, has an inner ring, an outer ring, a cage, and a single row of long and slightly crowned symmetrical rollers. As the name suggests, the rollers have toroidal shape. The raceways on both the inner and outer rings are concave and situated symmetrically about the bearing center. The design of raceway profiles with respect to the rollers permits the bearing to have the combined ability to accommodate both misalignment and axial displacement. The ability to accommodate misalignment is largely determined by the internal radial clearance. The internal radial clearance of the bearing can be accurately set by axially positioning the outer ring with respect to the inner ring. Toroidal roller bearings have a compact cross section normally seen in needle roller bearings. Due to a high degree of roller and raceway conformability, toroidal bearings have high radial load capacity and can handle minimal moment load.

Although toroidal roller bearings may have less frictional losses as compared with spherical roller bearings, sliding exists at contacts between rollers and raceways. Therefore, they may not be recommended for high-speed operation.



Self-Aligning Bearings, Fig. 4 A toroidal roller bearing

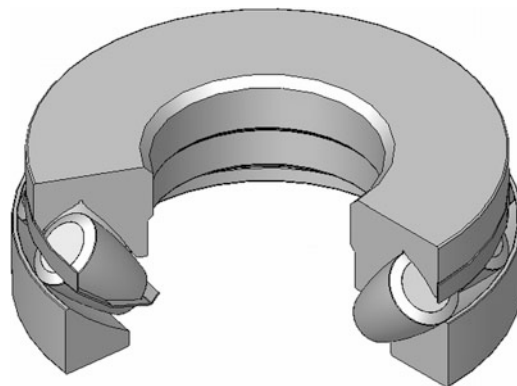
Applications of Toroidal Roller Bearings

Typical applications of toroidal roller bearings are seen in industrial planetary gear boxes, paper machines, and fans.

Self-Aligning Thrust Roller Bearing

Self-aligning thrust roller bearings are often referred to as spherical roller thrust bearings or tapered spherical roller bearings. They come in a single row arrangement, having two race rings and a set of rollers sandwiched between the race rings. Most bearings also have a cage to separate and retain the rollers on one of the raceways. [Figure 5](#) shows an example of tapered spherical roller bearings. In this bearing one of the raceways is constructed as a portion of a sphere to facilitate internal dynamic self-alignment; the rollers are asymmetrical and have spherical roller ends. The bearing has a concave and precisely ground rib that guides the spherical ends of the rollers, providing seating force to prevent rollers from being expelled out of the bearing raceways. Sliding occurs at the contacts between the roller ends and the guiding rib and at contacts between the roller bodies and raceways. Like spherical roller radial bearings, spherical roller thrust bearings are not suitable for high-speed operation.

Self-aligning thrust roller bearings have very high load-carrying capacity due to the close conformity between rollers and raceways. They can carry a combined thrust and radial load while permitting certain amount of misalignment between the housing bore and bearing shaft.



Self-Aligning Bearings, Fig. 5 A self-aligning thrust tapered roller bearing (courtesy of the Timken Company)

Application of Spherical Thrust Roller Bearings

Examples of applications for self-aligning thrust roller bearings are seen in wind turbines and heavy industrial structures.

Cross-References

- ▶ [Function and Structure of Rolling Element Bearings](#)
- ▶ [Multi-row Bearings](#)
- ▶ [Radial Bearings](#)
- ▶ [Rolling Element Bearings, History](#)
- ▶ [Thrust Bearings](#)

References

- B.J. Hamrock, *Ball Bearing Lubrication, The Elastohydrodynamics of Elliptical Contacts* (Wiley, New York, 1981)
- T.A. Harris, *Rolling Bearing Analysis*, 4th edn. (Wiley, New York, 2001)

Self-Alignment Bearings

- ▶ [Self-Aligning Bearings](#)

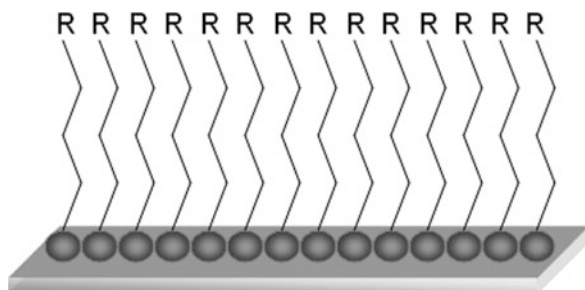
Self-assembled Monolayers

YUHONG LIU

State Key Laboratory of Tribology, Tsinghua University, Beijing, People's Republic of China

Definition

Self-assembled monolayers (SAMs) are two-dimensional molecular films organized by one layer of amphiphilic



Self-assembled Monolayers, Fig. 1 Scheme of self-assembled monolayer

molecules assembled on the substrate by a special affinity (Fig. 1). The thickness of the SAMs is usually between 3 and 10 nm and depends on the dimension of the molecular structure and the orientation angle of the molecule chain to the substrate surface. Self-assembled monolayers can be prepared using different types of molecules and different substrates. A classical example of SAMs is the alkanethiols molecules on a gold surface and functionalized silane compounds self-assembled on silicon surfaces, which have been investigated in great detail. Generally, they were formed simply by immersing a substrate into the solution or exposing it to the vapor phase of the desired molecule and incubating for a while. During the assembling process, the hydrophilic “head groups” are absorbed onto a substrate from either the vapor or liquid phase followed by a slow, two-dimensional organization of hydrophobic “tail groups.” At the beginning, amphiphilic molecules form either a disordered mass of molecules or a “lying down phase” (Fig. 2). Then after a few hours, crystalline or semicrystalline structures gradually begin to form on the substrate surface. Rather than a technique such as chemical vapor deposition or molecular beam epitaxy to add molecules to a surface (often with poor control over the thickness of the molecular layer), the self-assembled procedure is convenient and the SAMs are bonded in an orderly and compact manner with the substrate.

Scientific Fundamentals

It is a universal phenomenon in nature that molecules form into orderly nano-structures by self-assembly, which refers to the process wherein the molecules combine spontaneously by noncovalent bonds to form stable molecular aggregates in equilibrium condition. The force of noncovalent bonds can be the weak interactions of electrostatic, hydrogen, and coordination bonds, and others. This phenomenon where molecules form a specific shape by the given combinative manner is called

“self-assembly” (Ulman et al. 1991). Self-assembled monolayers refer to long chain molecules adsorbing on the surface of the proper substrate and arranging spontaneously by noncovalent bonds to form the orderly structure (Tripp and Hair 1995). The self-assembly technique is one kind of adsorption based on chemical reaction. It is very useful in forming monolayer or multiple layers of ultrathin molecular films with good physical or chemical characteristics. Hence, the self-assembly method has an implicit advantage in nanoscale-ordered membrane technology. The preparation of self-assembled monolayers is very simple and the influence of substrate shape is insignificant. Self-assembled monolayers are very useful for property modification on solid surfaces and have wide application in nanomanufacturing.

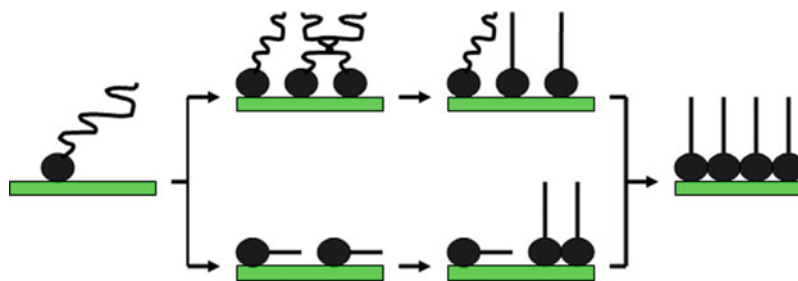
Self-assembly is very useful for surface modification. Orderly, compact, and stable two- or three-dimensional ultra-thin molecular films can be conveniently obtained by introducing the self-assembly technique. Due to the importance of self-assembled monolayers in technology research and their wide application in industry, SAMs have been widely studied and various preparation methods have been proposed.

Chemical Adsorption

Chemical adsorption utilizes chemical bonds to form molecular films. This method can enhance the mechanical or thermodynamic performance of the self-assembled monolayers. However, in order to guarantee the successful self-assembly of multiple layers of molecular films, various organic molecules that have different functional groups should be introduced onto the substrate. Furthermore, the multilayer films should be orderly and in periodicity; hence, the functional groups of the upper layer should be reacted entirely with the underlayer, or else structural defects would be caused. The disadvantage of chemical adsorption is that the repetition of experimental results is poor (Kim and Sohn 2001).

Molecule Aggradation

Molecule aggradation technique was first proposed by Decher (1997) in 1991. The principle of molecule aggradation is simple: a solid surface with positive charge is immersed into a solution with negative charge and incubated for a while. The solid surface is rinsed with distilled water after adsorption, and the surface is negatively charged. Then, the solid surface is immersed into the solution with positive charge and the surface is positively charged. Repeating this process, multilayer self-assembled molecular films are prepared by the weak electrostatic interactions. This orderly film aggradation method is



Self-assembled Monolayers, Fig. 2 Scheme of self-assembly procedure

a balanceable process with the following advantages: layers are coalesced with electrovalent bonds, preparation is simple, films are stable, and the influence of the shape or size of the substrate is insignificant. In addition, it is convenient to control the thickness of the self-assembled molecular layers and the structure of the multilayers at nanoscale. J. McCarthy et al. have utilized the molecule aggradation technique to study the application of polyethylene terephthalate (PET) to surface self-assembly modification. Because the adjacent layers are connected through electrostatic interactions, the above-mentioned method is also called electrostatic self-assembly. The thickness, structure, and chemical composition can be adjusted easily at the molecular scale; hence, the surface property that has an important influence on the performance of photoelectric devices can also be controlled. In addition, the substrates are not confined to smooth ones, and molecular films can also be self-assembled on anomalous surfaces.

Spin-Coating

The process of spin-coating is simple. A polymer solution is dripped onto a high-speed rotating substrate and the films are prepared. The formation of the special structure of the self-assembled monolayers is based on the mutual interactions between polymer molecules, and the thickness of SAMs could be controlled by changing the concentration of the polymer solution or the revolving speed of the substrate. Kim et al. used the spin-coating method and successfully self-assembled unsymmetrical polydiacetylene SAMs onto the glass substrate. With the help of Fourier transform infrared spectroscopy (FT-IR) and dielectric spectrum, they confirmed the hydrogen bond reticulate structure of self-assembled monolayers. Zhang et al. used glass as the substrate and obtained the second-order nonlinear optical self-assembled monolayers. With infrared spectrum, atomic force microscopy (AFM), and differential scanning calorimetry (DSC), they studied the formation process of

the self-assembly structure. Chen et al. also used the spin-coating technique and prepared light-emitting diode (LED) polymer SAMs.

Ingraft

With immersion of the ingrafted polymer at the solid surface into the solvent, if the polymer is not dissolved in the solvent and the ingrafting is not uniform, the macromolecular chains would be self-assembled to orderly films due to the solvophobic interactions. Balazs et al. have used the ingraft method and obtained homopolymer and copolymer self-assembled monolayers. The ingraft technique has many advantages: the size of the film can be controlled by changing the length of polymer chains, adjusting the property of solvent, or changing the ingrafting density.

The characteristics of self-assembled monolayers should be analyzed in detail after the films are prepared. With developments in science and technology, many instruments have been fabricated to investigate the characteristics of SAMs. The most popular approach is electrochemistry analysis, which has many attributes, such as high sensitivity, convenience, simple manipulation, and low cost. Almost all film systems can be detected by electrochemistry measurement. X-ray photoelectron spectroscopy (XPS) is used extensively as a surface analysis technique, and it is very useful for revealing the chemical bond properties of bonding atoms bonded with the substrate. Structure or chemical composition of a solid surface can be uncovered clearly by XPS measurement, hence the specific information of self-assembled monolayers can be determined. Furthermore, XPS can also be used for the quantitative analysis of an element. The spectral intensity of photoelectrons reflects the concentration or content of an element. The structures of SAMs are most commonly determined using scanning probe microscopy techniques such as atomic force microscopy (AFM) and scanning tunneling microscopy (STM). More recently, however, diffractive methods have also been used.

The structure can be used to characterize the kinetics and defects found on the monolayer surface. These techniques have also shown physical differences between SAMs with planar substrates and nanoparticle substrates. It is important to study the interactions between molecule and molecule and molecule and substrate. This can help discover the essence of the ordered growth, adhesion, and lubrication of self-assembled monolayers. A variety of other self-assembled monolayers can be formed, although there is always debate about the degree to which systems self-assemble.

Alkyl thiols are known to assemble on many metals, including silver, copper, palladium, and platinum. Alkyl silane molecules (e.g., octadecyltrichlorosilane) are well-known for self-assembly on silicon oxide surfaces and are potentially of greater technical relevance than alkyl thiol assembly on metals. Alkyl carboxylates are known to assemble on a variety of surfaces, such as aluminium and mica. Silicon has been used through the reaction of a silicon hydride surface and a radical generator, such as heat, UV, or radical initiator molecules, or with reagents such as Grignard and chlorosilanes. Once assembly has been accomplished, chemistry can be performed on the layer, especially if self-assembly places a reactive functional group on the outside of the monolayer. Some commonly used SAMs include 8-amino-1-octanethiol, hydrochloride, 6-amino-1-hexanethiol, hydrochloride, 10-carboxy-1-decanethiol, and 7-carboxy-1-heptanethiol.

One kind of self-assembled monolayer is alkylsilane SAMs deposited on silicon substrate. It is well known that silicon (Si) is the most popular material in MEMS components, but Si demonstrates very poor tribological performance (Satyanarayana and Sinha 2005). Alkylsilane self-assembled monolayers have been extensively studied and proposed as the lubricants for MEMS (Bhushan et al. 1995). Carbon chain length of the alkylsilane has an influence on its characteristics. The tribological properties of perfluoro alkylsilane SAMs are different from the non-perfluoro ones. It has been found that alkylsilane molecular films with shorter carbon chains exhibited a higher friction coefficient than the ones with long carbon chains due to a more disordered film structure (Kasai et al. 2005). Christian et al. (Lorenz et al. 2005) showed that the friction coefficient for the hydrocarbon alkylsilane self-assembled monolayers was approximately the same as for the fluorocarbon self-assembled monolayers. Kim et al. (1997) revealed that a threefold increase in friction could be observed upon the introduction of fluorine, and they attributed such an increase to the relatively large size of the fluorine compared with the hydrogen in the terminal group. Hoque et al. (2007) showed that once

self-assembled on Al, the perfluorinated alkyl chains were more stable than the simple hydrocarbon chains. Linford and Chidsey confirmed that alkyl could be covalently bonded to the silicon substrate, forming rigid self-assembled monolayers. Organic alkylsilane self-assembled monolayers are highly ordered and directional. The head or terminal functional groups of alkyl chain can have chemical reactions with many other functional groups, and the properties of the alkyl SAMs are dependent on the terminal groups of the alkyl chain. Hence, various characteristics could be obtained by modifying the surface with different functional terminated alkylsilane groups. It is true that alkylsilane self-assembled monolayers have wide application in making new functional materials and nano-devices. The substrates for forming the self-assembled molecular films of organic chlorosilane, hydrochloride, or amino-hydrochloride should be hydroxylated. The organic alkylsilane SAMs could be self-assembled on many substrates such as SiO₂, Al₂O₃, quartz, glass, mica, ZnSe, GeO, and Au. The intrinsic formation process of alkylsilane SAMs is as follows: at the beginning, alkyl chains are adsorbed to the hydroxylated silicon surface through the polar head group. When the group approaches the surface, hydrolyzation will be induced and alkyl chains will be connected with adjacent molecules and the surface hydroxy alkylsilane through hydrogen bond. However, this situation is not a stable state; the shrinking reaction will occur, and, as a result, all molecular chains will be connected as reticulation, and stable and uniform-oriented, self-assembled monolayers are formed.

Another example is alkyl thiol on gold. Sulfur has particular affinity for gold with a binding energy on the order of 45 kcal/mol (Dubois and Nuzzo 1992). An alkane with a thiol head group will stick to the gold surface and form an ordered assembly with the alkyl chains packing together due to van der Waals forces. For alkyl thiols on gold, the extended alkyl chains typically orient with an angle of $\sim 30^\circ$ from the perpendicular of the substrate (Porter et al. 1987), and are assumed to be in a fully extended linear arrangement. There has been a great deal of work done determining the process by which alkyl thiol on gold assemblies are produced. It is generally thought that alkyl thiol molecules first bind to the gold surface in a "lying down" position, where the alkyl chain tails of the molecules lie flat on the gold surface. The thiol interaction provides about 45 kcal/mol of driving force for the initial binding (Boeckl and Graham 2006), which is modeled as a Langmuir binding isotherm. These binding events continue until the lying down molecules are dense enough on the surface to interact with each other. At some point, the

alkyl chains lift off the substrate and point outwards, tethered by the thiol anchor to the surface. There is a shift to a mixture of lying down molecules and island domains of upright alkyl chains, tilted at 30° to normal. At this stage, binding kinetics become more complex and can no longer be modeled with a simple Langmuir binding isotherm. Over time, the island domains merge and cover the bulk of the substrate, and the process can be compared to a 2-D crystallization process on a surface. Alkyl thiol SAMs exhibit grain boundaries and defects even after long periods of assembly. The initial stage of SAM formation usually takes minutes or less under the normal conditions of 0.1–10 mmol/L thiol concentration in a solvent. More ordering of the assembly can take place over days or months, depending on the molecules involved.

Key Applications

In reality, self-assembled monolayers are usually used as the model to study the wettability, adhesion (Nakagawa and Ogawa 1994), lubrication, and corruption of various systems or the adsorption of protein and carbon nanotube (Xu and Hu 1995). The investigation of self-assembled monolayers is a hotspot in nano-technology, and SAMs have been used widely in hard disk, micro-electromechanical systems, and nano-electromechanical systems (MEMS/NEMS). Many research institutes place their interests on the investigation of alkane thiol on gold and alkylsilane self-assembled on hydrophilic substrates. SAMs can serve as models for studying membrane properties of cells and organelles and cell attachment on surfaces. SAMs can also be used to modify the surface properties of electrodes for electrochemistry, general electronics, and various micro- or nano-devices. For example, the properties of SAMs can be used to control electron transfer in electrochemistry. They can serve to protect metals from harsh chemicals and etchants. SAMs can also reduce sticking of MEMS and NEMS components in humid environments. In the same way, SAMs can alter the properties of glass. A common household product, Rain-X, utilizes SAMs to create a hydrophobic monolayer on car windshields to keep them clear of rain. SAMs have several other applications in scientific research. They tend to have quite different chemical kinetics, when the same molecules are in different forms, owing to their exposed, two-dimensional distribution, and as such are useful for some chemical and biochemical experiments. They can also be used for simulation of biological membranes and as substrates for cell culture. As technology develops to control the functional groups in SAMs, either by direct deposition of molecules with these groups or by chemical modification of the layer, it allows many other

applications, to develop, for example, the fabrication of nanoscale electronics.

Cross-References

- [Polymer Nanolayers](#)
- [Solid-Like Lubricating Films, Ionic Liquid Films](#)

References

- B. Bhushan, A.V. Kulkarni, V.N. Koinkar et al., *Langmuir* **11**, 3189 (1995)
- M. Boeckl, D. Graham, *Mater. Matters* **2**, 15 (2006)
- C.Z. Decher, *Science* **277**, 1232 (1997)
- L.H. Dubois, R.G. Nuzzo, *Annu. Rev. Phys. Chem.* **43**, 437 (1992)
- E. Hoque, J.A. DeRose, P. Hoffmann et al., *J. Phys. Chem. C* **111**, 3956 (2007)
- T. Kasai, B. Bhushan, G. Kulik et al., *J. Vac. Sci. Technol. B* **23**, 995 (2005)
- T.K. Kim, B. Sohn, *Appl. Surf. Sci.* **244**, 109 (2001)
- H.I. Kim, T. Koini, T.R. Lee et al., *Langmuir* **13**, 7192 (1997)
- C.D. Lorenz, M. Chandross, G.S. Grest et al., *Langmuir* **21**, 11744 (2005)
- T. Nakagawa, K. Ogawa, *Langmuir* **10**, 367 (1994)
- M.D. Porter, T.B. Bright, D.L. Allara, C.E.D.J. Chidsey, *Am. Chem. Soc.* **109**, 3559 (1987)
- N. Satyanarayana, S.K.J. Sinha, *Phys. D: Appl. Phys.* **38**, 3512 (2005)
- C.P. Tripp, M.L. Hair, *Langmuir* **11**, 149 (1995)
- A. Ulman, (Boston, Academic Press, 1991)
- J. Xu, J. Hu, *J. Colloid Interface Sci.* **176**, 138 (1995)

Self-Assembled Monolayers for Boundary Lubrication

- [Chemical Vapor Deposition Processes for Boundary Lubricants](#)

Self-Excited Gas Bearing Instabilities

CODA H. PAN
Global Technology, Millbury, MA, USA

Synonyms

[Axial instability](#); [Conical whirl](#); [Fractional frequency whirl](#); [Half frequency whirl](#); [Non-synchronous whirl](#); [Resonant whip](#); [Self-excited shaft whirl](#)

Definition

A mechanical system that is lubricated with either liquid or gas film bearings can experience an anomalous state of

self-excited instability. This essay is focused on instability of gas film bearings (► [Hydrodynamic Journal Bearings](#)).

Scientific Fundamentals

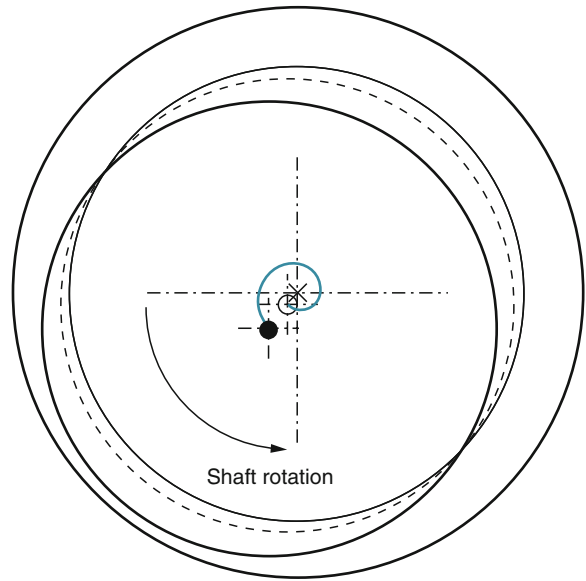
Among various types of gas bearing instabilities, the unstable whirl is most notorious. It is characterized by a growing orbital motion of the shaft center(s). The orbital frequency is typically about half of the shaft rotation rate. Once initiated, the orbit amplitude would increase incessantly in the form of an outwardly spiraling trajectory until sliding contact occurs to cause catastrophic failure of the bearing(s). As illustrated in Fig. 1, an unstable whirl orbit is triggered by a small initial displacement of the shaft center at \circ , and follows an outward spiraling path, which is depicted in blue, to reach a near-contact location at \bullet after approximately two shaft rotations.

When simultaneously viewed in two mutually perpendicular radial planes, the self-amplifying whirl motion is seen as a pair of displacements-versus-time traces that are phased in quadrature with an exponential increasing envelop as illustrated in Fig. 2; for the particular instability growth rate, touch-down takes place at \times in a little over two rotations.

Gas lubricated thrust and conical bearings are also susceptible to instability in the form of low frequency oscillations. This mode of instability was commonly associated with externally pressurized gas bearings, known as “pneumatic hammer” (► [Pneumatic Hammer](#)); it can also take place in spiral groove textured bearings (► [Gas Bearings with Narrow Inclined Grooves](#)).

In general, a rotor can interact with the supporting bearing system in five degrees of freedom in the form of displacements from the equilibrium rotor center in the axial direction (δz), parallel translation of the rotor axis in two mutually perpendicular radial directions (δx , δy) and angular displacement of the rotor axis in two mutually perpendicular radial planes ($\delta \xi$, $\delta \eta$) as illustrated in Fig. 3.

Components of translational displacement are coupled with quadrature phasing to form cylindrical whirl, which is the same at either end of the rotor axis as illustrated in Figs. 1 and 2. Similarly coupled with quadrature phasing while the two ends are oppositely displaced from the rotor axis is the conical whirl; upon displayed together, will appear similarly as Fig. 2, but in superimposed mirror images. Hence the five degrees of freedom are combined into three modes, namely axial (displacement), cylindrical (whirl), and conical (whirl) modes for idealized rotor systems that possess symmetrical properties. In real systems these modes are often coupled with one another.



Self-Excited Gas Bearing Instabilities, Fig. 1 Trajectory of unstable whirl

Prevention of Instabilities

The happenstance of gas bearing instability on an existing hardware is a very costly experience. Engineering study to assure adequate margin of safety from instability is an essential preparation in consideration of a deployment of gas bearings.

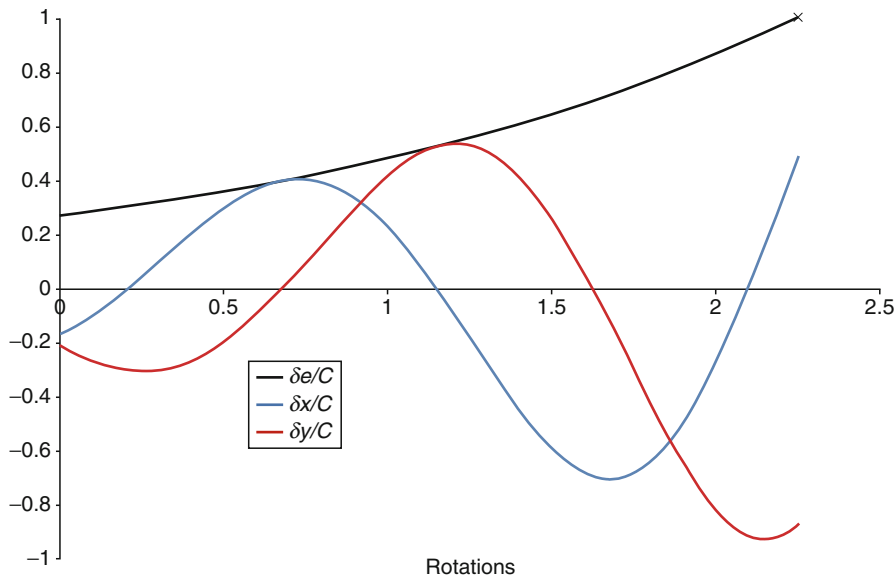
The most expedient option to avoid instabilities is to select a bearing design with established application track records of field reliability; examples comprise tilting-pads bearings (► [Tilting Pad Gas Bearings](#)), foil bearings (► [Foil Gas Bearings](#)), spiral-grooved gas bearings (► [Gas Bearings with Narrow Inclined Grooves](#)).

In light-duty motor driven applications, a viable approach is to combine inherent magnetic loading with a relatively long plain journal bearing. Active magnetic damping is a possibility if hardware requirements can be accommodated.

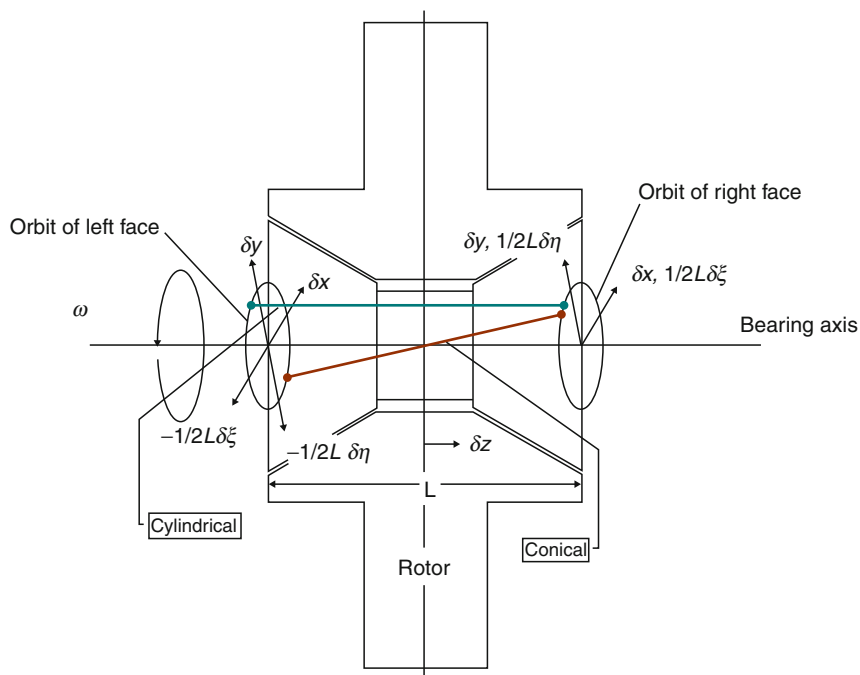
Alternative Approaches

Having identified the preferred design approach, risk of instability can be assessed either by dynamic simulation or by performing stability analysis.

Dynamic simulation – In dynamic simulation, a steady operating environment is specified together with an assumed initial state, e.g., the initial state of the rotor axis, then trajectory(ies) of rotor center(s) is computed according to the condition of dynamic equilibrium



Self-Excited Gas Bearing Instabilities, Fig. 2 Whirl motion seen as displacement-versus-time traces



Self-Excited Gas Bearing Instabilities, Fig. 3 Five degrees of freedom of rigid rotor

between the rotor body and the resultant gas film force; the latter is determined by computing the time-dependent system response according to the governing equations as applicable to the selected bearing design. A satisfactory

design is indicated by attenuating amplitudes of rotor trajectories that settle down to a final steady state. Margin of safety of the design is represented by the decay rate of the attenuating trajectories.

Dynamic simulation relies entirely on numerical computation. Its validity depends on the accuracy of the computation procedure, especially because the time-dependent gas-film continuity equation has the parabolic structure, second order spatial derivatives concurrent with first order time derivative, it is vulnerable to numerical instability that can falsify the simulation process.

Dynamic simulation is routinely used to study air lubricated flyheads used in computer hard disk drives (► [Molecular Gas Film Lubrication](#)). Flyhead dynamics is concerned with three modes of motion; namely vertical, pitch and roll motions that are always coupled. Typical flyhead designs are inherently stable. Dynamic simulation is used mainly to find the steady-state flying height while sampling transient oscillations regarding amplitudes and decay time.

Stability analysis—Stability analysis begins by defining the time-independent solution of the gas film pressure that is in equilibrium with the specified operating environment. Dynamic perturbation from the established steady-state solution is performed with allowance of time-dependence; mathematical linearization is imposed here by allowing the perturbation amplitude to be arbitrarily small without restricting the character of time-dependence. Dynamic equilibrium with the rotor body is examined upon separating out the time-independent part; the condition of a self-sustained transient motion is accordingly determined, the latter includes a set of stability parameters that represent a comprehensive characterization of the rotor-bearing system. Typically a stability analysis yields a map that identifies domain(s) for safe operation of the rotor-bearing system.

Performance of stability analysis is a more demanding undertaking than dynamic simulation. A stability study would consider wide ranges of many parameters and make possible looking ahead when there is no experience or precedent to guide dynamic simulation.

Stability Analysis of Gas Bearing Supported Mechanical System

Physical Concept

Stability is concerned with three possibilities of the state of equilibrium; namely a stable state, a neutrally stable state, and an unstable state.

The possibility of the three states may be illustrated by the equilibrium of a mass point that is restrained by a spring-dashpot connection attached to the ground, with stiffness and damping coefficients (k, c). A steady force F is

exerted on the mass point to maintain a spring displacement x_0 ; a static equilibrium condition is satisfied by

$$F - kx_0 = 0$$

Note

- The damping coefficient c does not affect the static equilibrium state.
- The exerted force F is proportional to x_0 while $k > 0$ is a pre-requisite for the presence of the equilibrium state.

The system of interest encompasses

$$m; k, c \text{ and } F$$

To assess the stability issue, the dynamic equilibrium condition is to be examined; that is

$$F - kx - c\dot{x} - m\ddot{x} = 0$$

It is assumed that there is an infinitesimal deviation from the static equilibrium position:

$$\delta x = x - x_0$$

Subtracting out the static equilibrium condition, one obtains the perturbed differential equation of motion

$$-m\delta\ddot{x} = k\delta x + c\delta\dot{x}$$

This is a homogeneous linear equation of δx ; an exponential time-dependence can be assumed and factored out, thus

$$\delta x = \tilde{x} \exp\{st\}$$

Whereupon the perturbed differential equation of motion is transformed into the characteristic equation, which happens to be a second order polynomial:

$$(ms^2 + cs + k)\tilde{x} = 0$$

The exponential time constant can now be found as the quadratic root

$$s = \frac{-c \pm \sqrt{c^2 - 4mk}}{2m}$$

Consider various possible combinations of (k, c).

1. $k = 0$ $F = kx_0 \rightarrow 0$ for all finite x_0 . The characteristic roots are

$$s = (0, -c/m)$$

The first root identifies a neutral state of stability. The second one requires $c > 0$ to ensure dynamic stability since m is always positive.

2. $k < 0$ $\sqrt{c^2 - 4mk} > 0$, hence one of the roots has a positive real part to indicate a state of instability.

Any $\delta x \neq 0$ would be amplified exponentially without an oscillatory component. This is a *statically unstable state*.

3. $k > 0$ The real part of either root shares the same sign with c . Hence a stable state depends on $c > 0$. If $0 < c^2 < 4mk$, then writing $4m\Omega'^2 = 4mk - c^2$, the characteristic roots are complex conjugates

$$s = -\frac{1}{2}(c/m) \pm j\Omega'$$

$\frac{1}{2}(c/m)$ is the decay rate and $\Omega' < \sqrt{k/m}$ is the damped natural frequency. If $c^2 > 4mk$, both roots are without an imaginary part; there are two stable states with independent decay rates.

The restrained mass point example presents completely the physical concepts involved in the stability analysis of gas bearing supported systems.

General Scheme for Stability Analysis of Gas Bearing Support Systems

Although the restrained mass point example adequately describes all fundamental ingredients of perturbed transients, the effort required to carry out a complete stability analysis of the gas bearing support is substantially more demanding; it requires attention to additional issues as enumerated below:

1. All gas bearing support systems have several degrees of freedom with cross-coupling.
2. Coupled bearing impedance coefficients do not directly reveal the physical functions of stiffness and damping; the latter are derived from the roots of the characteristic equation of the impedance matrix that also stipulates the natural amplitude-weighting and phasing of the coupled degrees-of-freedom to form "self-sustained" natural motions.
3. Compressibility of the gas film calls for inclusion of time-derivative of perturbed pressure in the squeeze film effect.
4. Difficult numerical computation issues that require attention.

These questions have been extensively examined and can be considered to be fundamentally resolved. Essential ingredients required of a valid stability analysis of gas bearing supports are outlined here.

Treatment of the Bearing Film Pressure

The support capability of a gas bearing results from the film pressure p , which obeys the lubrication theory and is adequately modeled as an isothermal film of a perfect gas.

Due to viscosity, the gas film tends to attach to the enclosing surfaces. In a rarefied state, however, with allowance for molecular effects, slip occurs at both enclosing walls so that a gas film, which is contained between closely spaced impermeable bearing surfaces, obeys the following formulas, Burgdorfer (1959), Kang et al. (1997), and Li (2002):

$$\begin{aligned}\bar{\tau} &= \sigma_\tau \mu \Delta \bar{V} / h \\ p\bar{\phi} &= \sigma_C \bar{V} ph - \sigma_P (12\mu)^{-1} ph^3 \text{grad } p\end{aligned}\quad (1)$$

Slip effects do not affect gas film related stability problems aside from modifying the governing equations for shear and pressure as indicated above. The continuum approximation will henceforth be assumed, so that with $(\sigma_\tau, \sigma_C, \sigma_P) = 1$, the gas film is governed by

$$\begin{aligned}\bar{\tau} &= \mu \Delta \bar{V} / h \\ p\bar{\phi} &= \bar{V} ph - (12\mu)^{-1} ph^3 \text{grad } p \\ \text{div}\left(p\bar{\phi}\right) + \frac{\partial(ph)}{\partial t} &= 0\end{aligned}\quad (2)$$

This is a mathematical statement of a two-dimensional field of the film pressure p with time dependence and is driven by the entrainment velocity \bar{V} . For steady-state operation, the last term on the left hand side is discarded, the steady-state film thickness function h_0 is specified, and the steady-state film pressure p_0 is solved from

$$\begin{aligned}p_0\bar{\phi}_0 &= \bar{V} p_0 h_0 - (12\mu)^{-1} p_0 h_0^3 \text{grad } p_0 \\ \text{div}\left(p_0\bar{\phi}_0\right) &= 0\end{aligned}\quad (3)$$

Projected integration of $\bar{v}(p_0 - p_a)dS$ renders the components of the steady-state bearing force:

$$(F_x, F_y, F_z) = - \iint_{\text{plan form}} \left(\bar{i}, \bar{j}, \bar{k} \right) \cdot \bar{v}(p_0 - p_a) dS \quad (4)$$

Integration of its triple scalar product with the transverse coordinates yields steady-state lateral moments:

$$(M_\xi, M_\eta) = \iint_{\text{plan form}} \left[\left(-\bar{j}, \bar{i} \right) \cdot \bar{v} \times \bar{k} z \right] (p_0 - p_a) dS \quad (5)$$

Deviation from the steady-state is defined by an infinitesimal displacement of the rotor; it can be stated in each of the five *DOFs* of a rigid rotor:

$$\begin{array}{ccc} \text{Axial} & \text{Traslational} & \text{Angular} \\ \delta z & \delta x, \delta y & \delta \xi, \delta \eta \end{array} \quad (6)$$

The Jacobian notation can be used to designate perturbation of the film thickness in each *DOF*

$$\begin{aligned}\delta h \equiv h - h_0 &= \frac{\partial h}{\partial(\delta x, \delta y, \delta z, \delta \xi, \delta \eta)} \\ &= h_x \delta x + h_y \delta y + h_z \delta z + h_\xi \delta \xi + h_\eta \delta \eta\end{aligned}\quad (7)$$

Film pressure perturbation due to δh is postulated as

$$\delta p \equiv p - p_0 = p_h \delta h \quad (8)$$

Substitution of (7) and (8) into (2), omitting high order terms subtracting out (3), it is found

$$\text{div} \delta(p\bar{\phi}) + \frac{\partial(p_0 \delta h + \delta p \delta h_0)}{\partial t} = 0 \quad (9)$$

Upon finding δp , its projected integration would render the perturbed bearing reactions. For the axial *DOF*

$$\delta F_z / \delta z = - \iint_{\text{plan form}} \bar{k} \cdot \bar{v} p_h h_z dS \quad (10)$$

Bearing perturbation reaction to the translational *DOF* may not be collinear; hence the matrix representation is used for the four components

$$\begin{aligned}-(Z_{xx}, Z_{yx}) &= (\partial \delta F_x / \partial \delta x, \partial \delta F_y / \partial \delta x) \\ -(Z_{xy}, Z_{yy}) &= (\partial \delta F_x / \partial \delta y, \partial \delta F_y / \partial \delta y)\end{aligned}\quad (11)$$

They are

$$\begin{aligned}(Z_{xx}, Z_{yx}) &= \iint_{\text{plan form}} \begin{pmatrix} \bar{i} & \bar{j} \end{pmatrix} \cdot \bar{v} p_h h_x dS \\ (Z_{xy}, Z_{yy}) &= \iint_{\text{plan form}} \begin{pmatrix} \bar{i} & \bar{j} \end{pmatrix} \cdot \bar{v} p_h h_y dS\end{aligned}\quad (12)$$

For the angular *DOF*, the perturbation reaction moment also may not be co-planar; again it is convenient to use matrix notation, therefore

$$\begin{aligned}-(Z_{\xi\xi}, Z_{\eta\xi}) &= (\partial \delta M_\xi / \partial \delta \xi, \partial \delta M_\eta / \partial \delta \xi) \\ -(Z_{\xi\eta}, Z_{\eta\eta}) &= (\partial \delta M_\xi / \partial \delta \eta, \partial \delta M_\eta / \partial \delta \eta) \\ (Z_{\xi\xi}, Z_{\eta\xi}) &= \iint_{\text{plan form}} \left[\begin{pmatrix} -\bar{j} & \bar{i} \end{pmatrix} \cdot \bar{v} \times \bar{k} z \right] p_h h_\xi dS \\ (Z_{\xi\eta}, Z_{\eta\eta}) &= \iint_{\text{plan form}} \left[\begin{pmatrix} -\bar{j} & \bar{i} \end{pmatrix} \cdot \bar{v} \times \bar{k} z \right] p_h h_\eta dS\end{aligned}\quad (13)$$

Dynamic Equilibrium of Rotor-Bearing System

Perturbed dynamic equilibrium pairs the d'Alembert force/moment with the perturbed bearing reaction force/moment. For the axial and translational *DOFs*

$$(\mathcal{D}_z; \mathcal{D}_x, \mathcal{D}_y) = -m(\delta \ddot{z}; \delta \ddot{x}, \delta \ddot{y}) \quad (14)$$

For the angular *DOFs*, there are two mass moments of inertia, J_P and J_T , respectively, about the rotation and a transverse axes; and there is a precession reaction in the corresponding d'Alembert moment

$$(\bar{\mathcal{D}}_\xi, \bar{\mathcal{D}}_\eta) = -J_T(\bar{i}_\xi \delta \ddot{\xi}, \bar{i}_\eta \delta \ddot{\eta}) + \omega J_P(-\bar{i}_\xi \delta \dot{\eta}, \bar{i}_\eta \delta \dot{\xi}) \quad (15)$$

Equation (9), due to (8), describes a homogenous linear system in $(\delta p, \delta h)$, so that time-dependence can be represented by an exponential function, which can be factored out; thus

$$(\delta p, \delta h) = (\delta \tilde{p}, \delta \tilde{h}) \exp\{st\} \quad (16)$$

Consequently, one finds it useful to adopt the practice of control technology by employing the Laplace substitution formula so that

$$\frac{\partial(\delta p, \delta h)}{\partial t} = s(\delta p, \delta h) \quad (17)$$

$$\text{div} \delta(p\bar{\phi}) + s(p_0 \delta h + \delta p h_0) = 0$$

The above is applicable to either one of all three cases according to how the film thickness Jacobian is assigned.

For the axial *DOF*, $\delta h = h_z \delta z$, δp as solved from (17) would be integrated according to (10) to render the axial perturbation bearing reaction. The axial d'Alembert force is given by (14); its dynamic equilibrium with the bearing reaction force requires

$$\delta F_z - m s^2 \delta z = 0 \quad (18)$$

For the translational *DOF*, $\delta h = h_x \delta x + h_y \delta y$, δp as solved from (17) would be integrated according to (12) resulting in coupled perturbation bearing reaction forces that are commonly stated in matrix notation. The (x, y) d'Alembert forces are also given by (14). Dynamic equilibrium obeys

$$\begin{bmatrix} (Z_{xx} + m s^2) \delta x & Z_{xy} \delta y \\ Z_{yx} \delta x & (Z_{yy} + m s^2) \delta y \end{bmatrix} = 0 \quad (19)$$

For the angular *DOF*, $\delta h = h_\xi \delta \xi + h_\eta \delta \eta$, δp as solved from (17) would be integrated according to (13) to form perturbation bearing reaction moments. Perturbation d'Alembert moments with gyroscopic precessions can also be stated in matrix notation.

$$\begin{aligned}(\mathcal{D}_{\xi\xi}, \mathcal{D}_{\eta\xi}) &= (-J_T s^2, s \omega J_P) \\ (\mathcal{D}_{\xi\eta}, \mathcal{D}_{\eta\eta}) &= (-s \omega J_P - J_T s^2)\end{aligned}\quad (20)$$

Dynamic equilibrium then requires

$$\begin{bmatrix} (Z_{\xi\xi} + J_T s^2) \delta \xi & (Z_{\xi\eta} + s \omega J_P) \delta \eta \\ (Z_{\eta\xi} - s \omega J_P) \delta \xi & (Z_{\eta\eta} + J_T s^2) \delta \eta \end{bmatrix} = 0 \quad (21)$$

Equation (19) and (21) are homogeneous equations defining a self-sustaining motion that calls for vanishing of the

determinant of the applicable matrix. Stability analysis in control technology has led to the development of stability assessment based on the polynomial type characteristic determinant; they are not usable here because a general solution of (9) does lead to such treatment; it is necessary to rely on its numerical treatment as an integral part of the stability analysis; it calls for an evaluation of the sign of $\lambda \equiv \text{Re}\{s\}$ with specified $(m; J_T, J_P)$. An indirect approach, which is easier to implement, seeks the value of $\text{Im}\{s\}$ for a state of neutral stability along with a criterion for $(m; J_T, J_P)$ to ensure stability. Thus, one sets

$$s_{\text{neutral}} = j\Omega\omega \quad (22)$$

Equation (17) is replaced by

$$\text{div}\delta(p\bar{\phi}) + j\Omega\omega(p_0\delta h + \delta p\delta h_0) = 0 \quad (23)$$

Then solutions of (10), (12), and (13) would all be complex to represent temporal phasing.

Axial DOF Perturbed bearing reaction of the axial DOF is

$$\delta F_z/\delta z = -Z_{zz}(\Omega) = -U_z(\Omega) - jV_z(\Omega) \quad (24)$$

Equation (18) is reduced to

$$U_z - (m\omega^2)\Omega^2 + jV_z = 0 \quad (25)$$

Vanishing of $V_z(\Omega_{\text{neutral}})$, the imaginary part of (25) is a required condition for establishing a valid neutral state stability. In addition it is necessary to abide by the physical condition that $(m\omega^2)\Omega^2$ must be positive; therefore, there is an additional requirement and the necessary and sufficient conditions for a neutrally stable axial state are

$$\begin{aligned} V_z(\Omega_{\text{neutral}}) &= 0 \\ U_z(\Omega_{\text{neutral}}) &= (m_{\text{threshold}}\omega^2)\Omega_{\text{neutral}}^2 > 0 \end{aligned} \quad (26)$$

Both (25) and (26) must be satisfied to identify a state of neutral stability with the parameters $(\Omega_{\text{neutral}}, m_{\text{threshold}})$.

Translational DOF The neutrally stable state of the translational DOF is defined by substituting (22) into (19), resulting in a homogeneous matrix equation, which must have a nil determinant

$$\begin{vmatrix} Z_{xx} - (m\omega^2)\Omega^2 & Z_{xy} \\ Z_{yx} & Z_{yy} - (m\omega^2)\Omega^2 \end{vmatrix} = [Z_{xx} - (m\omega^2)\Omega^2][Z_{yy} - (m\omega^2)\Omega^2] - Z_{xy}Z_{yx} \quad (27)$$

This is a second order polynomial in $(m\omega^2)\Omega^2$; hence the quadratic formula can be used to find

$$(m\omega^2)\Omega^2 = \frac{(Z_{xx} + Z_{yy})}{2} \pm \sqrt{\frac{(Z_{xx} - Z_{yy})^2}{4} + Z_{xy}Z_{yx}} \quad (28)$$

This is a second order polynomial in $(m\omega^2)\Omega^2$; hence the quadratic formula can be used to find

$$(m\omega^2)\Omega^2 = \frac{(Z_{xx} + Z_{yy})}{2} \pm \sqrt{\frac{(Z_{xx} - Z_{yy})^2}{4} + Z_{xy}Z_{yx}} \quad (29)$$

Similar as (25) in the axial DOF, in the translational DOF, (29) furnishes the necessary and sufficient conditions as

$$\begin{aligned} \frac{(V_{xx} + V_{yy})}{2} \pm \text{Im}\left\{\sqrt{\frac{(Z_{xx} - Z_{yy})^2}{4} + Z_{xy}Z_{yx}}\right\} &= 0 \\ (m\omega^2)\Omega^2 = \frac{(U_{xx} + U_{yy})}{2} \pm \text{Re}\left\{\sqrt{\frac{(Z_{xx} - Z_{yy})^2}{4} + Z_{xy}Z_{yx}}\right\} &> 0 \end{aligned} \quad (30)$$

Angular DOF Mathematical aspects of the angular DOF stability analysis are very similar to those of the translational DOF. Hence, without further elaboration, one can set to zero the determinant of (21) and state its roots as

$$\begin{aligned} J_T(\Omega\omega)^2 &= \frac{1}{2}(Z_{\xi\xi} + Z_{\eta\eta}) \\ &\pm \sqrt{\frac{1}{4}(Z_{\xi\xi} - Z_{\eta\eta})^2 + (Z_{\xi\eta} + j\Omega J_P\omega^2)(Z_{\eta\xi} - j\Omega J_P\omega^2)} \end{aligned} \quad (31)$$

This leads to necessary and sufficient conditions for the neutral state of stability as

$$\begin{aligned} &\frac{1}{2}(V_{\xi\xi} + V_{\eta\eta}) \\ &= \mp \text{Im}\left\{\sqrt{\frac{1}{4}(Z_{\xi\xi} - Z_{\eta\eta})^2 + (Z_{\xi\eta} + j\Omega J_P\omega^2)(Z_{\eta\xi} - j\Omega J_P\omega^2)}\right\} \\ &J_T(\Omega\omega)^2 \\ &= \frac{1}{2}(U_{\xi\xi} + U_{\eta\eta}) \\ &\pm \text{Re}\left\{\sqrt{\frac{(Z_{\xi\xi} - Z_{\eta\eta})^2}{4} + (Z_{\xi\eta} + j\Omega J_P\omega^2)(Z_{\eta\xi} - j\Omega J_P\omega^2)}\right\} > 0 \end{aligned} \quad (32)$$

Stability Criterion A neutral state of stability is established by carrying out a full-frequency survey to test its existence according to (26), (30), or (32). Upon finding Ω_{neutral} that satisfies the necessary and sufficient existence conditions, the mass parameter $m_{\text{or}J_T}$ is designated as the threshold value for incipience of instability and is given an analytical variation (AV) operation in accordance with the applicable dynamic equilibrium condition.

The AV operation is concerned with the condition for incipience of instability that can be characterized by an infinitesimal departure by the Laplace coefficient from an entirely imaginary with a variation of the mass parameter $\delta m_{\text{or}J_T}$. The theory of complex variables for complex

analytical function is applied to the computed bearing perturbation reactions in making use of

$$d(\cdot)/ds = \partial(\cdot)/\partial(j\Omega\omega) \quad (33)$$

The computed AV parameter $\delta s/\delta m$ has both real and imaginary parts. The real part, $\delta\lambda/\delta m$, determines whether the threshold mass parameter is an upper- or a lower-bound for stability and sets the amplitude growth rate upon triggering instability. The imaginary part, $\delta\Omega/\delta m$, sets the associated frequency-shift.

Illustrative Examples

Existence of self-excited instability in a spiral-grooved thrust bearing was made known by Malanoski and Pan (1965). The neutral state of stability of the axial DOF is established by vanishing of $V_z(\Omega_{\text{neutral}})$ and equating $m_{\text{threshold}}$ to $U_z/(\Omega\omega)^2$. Full-frequency survey can be displayed in terms of an impedance contour plot in a complex plane, which is a graph of $V_z(\Omega)$ versus $U_z(\Omega)$ with Ω stepped from a nil value to a suitably large value to demonstrate completeness of the survey in that the asymptotic trend of (24) is in evidence:

$$\lim_{|\Omega| \rightarrow \infty} \delta F_z/\delta z = -U_{z,\infty} - j\Omega^{-1}V_{z,\infty} \quad (34)$$

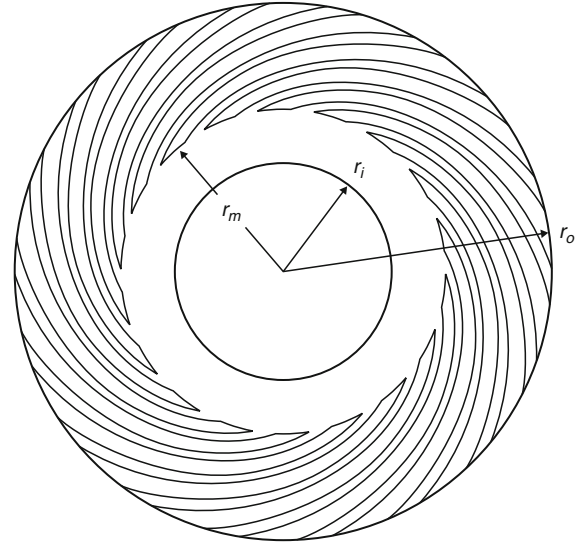
A compilation of axial perturbation impedance of an inward-pumping spiral-grooved thrust bearing (► [Gas Bearings with Narrow Inclined Grooves](#)) can be used to demonstrate the procedure for stability assessment in the axial DOF. The plan form of this bearing is shown in Fig. 4.

Perturbation impedance was computed for atmospheric ambient pressure at 50°C. In Fig. 5 are shown impedance contour graphs for $A = (10, 20)$. Smooth closure at to the real axis at the largest computed $|\Omega\omega| = 10^7 \text{ Hz}$ is seen in both cases; completeness of the compilation is ascertained.

Consider now the AV operation regarding axial dynamic equilibrium. An intercept of the real axis of the impedance contour, where V_z vanishes, existence of a neutrally stable state depends on the test

$$U_z|_{V_z=0} > 0? \quad (35)$$

All real axis intercepts pass the above test and can be identified as neutral stability points. An impedance contour of the axial DOF is constructed of mirror images of $\pm |\Omega|$, which are conjugate solutions of (24); while the numerical aspects of the conjugate solution is redundant, its existence serves to verify the completeness of analytic functions.



$$\begin{array}{llll} r_i = 8 \text{ mm} & R_m = 11.2 \text{ mm} & R_o = 20 \text{ mm} & C = 10 \text{ }\mu\text{m} \\ \alpha = 0.659 & & \beta = 18^\circ & \Gamma = 3.05 \end{array}$$

Self-Excited Gas Bearing Instabilities, Fig. 4 Plan form of inward-pumping thrust bearing

Upon postulating infinitesimal deviations δm and δs , one finds dynamic equilibrium in $O\{\delta\}$ is satisfied by

$$-(dZ_{zz}/ds)\delta s = \delta m s^2 + 2ms\delta s$$

Arrange into an equation for finding δs from δm , with all coefficients identified with the neutral state

$$\delta s = \frac{-s^2 \delta m}{(dZ_{zz}/ds) + 2ms} \Big|_{s_{\text{neutral}}}$$

Then compute the impedance derivative using the formula for analytic functions of a complex variable

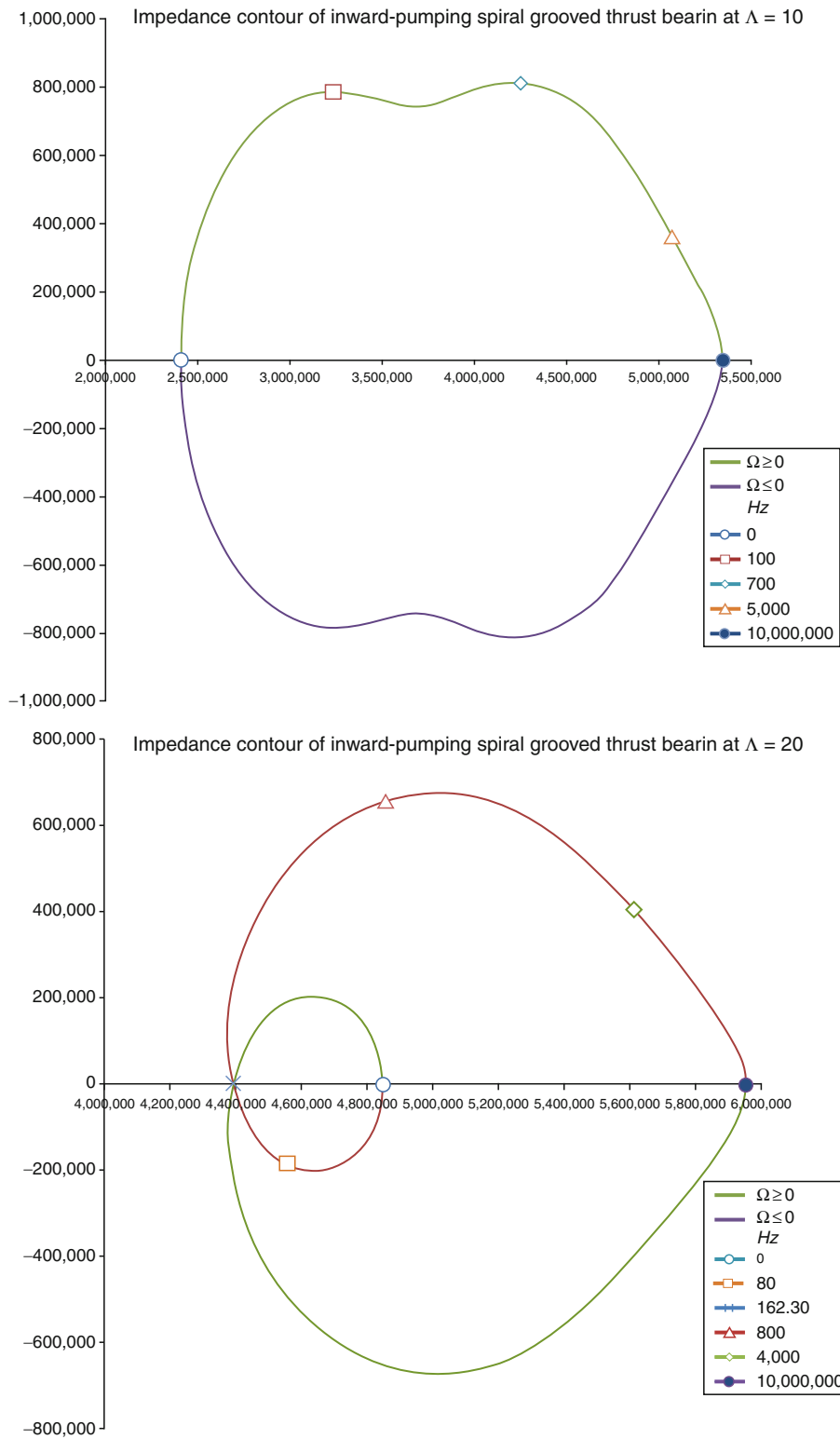
$$\delta s = \frac{-j[(\partial Z_{zz}/\partial \Omega) - 2m\Omega\omega]_{\text{conjugate}} s^2 \delta m}{|(\partial Z_{zz}/\partial \Omega) - 2m\Omega\omega|^2} \Big|_{s_{\text{neutral}}} \quad (36)$$

Separating real and imaginary parts, it is found

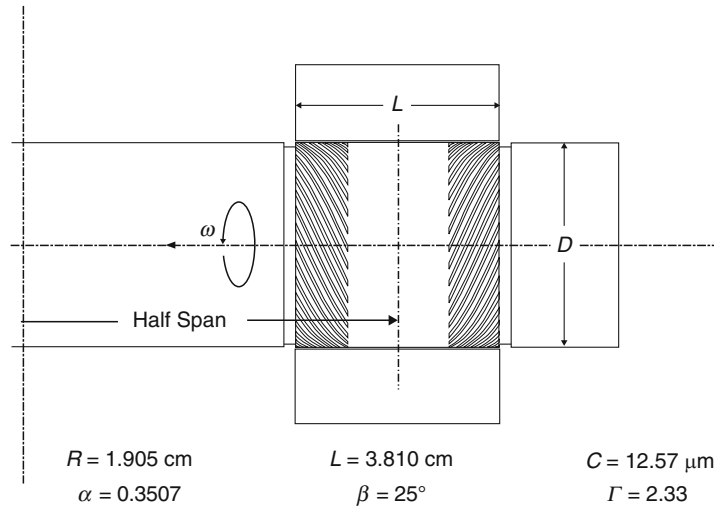
$$\delta\lambda = \frac{(\partial V_{zz}/\partial \Omega)(\Omega_{\text{neutral}}\omega)^2 \delta m}{|(\partial Z_{zz}/\partial \Omega) - 2m\Omega_{\text{neutral}}\omega|^2} \quad (37)$$

and

$$\delta(\Omega\omega) = \frac{(\Omega_{\text{neutral}}\omega)^4 (\partial/\partial \Omega) \{U_z/(\Omega_{\text{neutral}}\omega)^2\} \delta m}{|(\partial Z_{zz}/\partial \Omega) - 2m\Omega_{\text{neutral}}\omega|^2} \quad (38)$$



Self-Excited Gas Bearing Instabilities, Fig. 5 Impedance contours of inward-pumping spiral-grooved thrust bearing



Self-Excited Gas Bearing Instabilities, Fig. 6 Herringbone journal bearing on a shaft

The above is an algebraic proclamation of the Stability Criteria for the axial *DOF*. A verbal statement of the stability theorem follows:

- With an infinitesimal increment of the rotor mass beyond its threshold value, the system would be unstable if and only if the imaginary part of the bearing reaction impedance $V_z(\Omega)$ attains an algebraic increment as Ω exceeds Ω_{neutral} . Conversely, if the imaginary part of the bearing reaction impedance attains an algebraic decrement with as Ω exceeds Ω_{neutral} , an infinitesimal increment of the rotor mass beyond its threshold value would cause the system to become stable.

Returning to the impedance contours, the above stated stability theorem can be used to examine all intercepts on the real axis. For $A = 10$, the intercept at $\Omega = 0$ is an upward crossing line, hence according to (37), the threshold mass is an upper bound for stability; however its magnitude is unbounded, hence it poses no stability problem for rotors of finite mass. The intercept at $|\Omega| \rightarrow \infty$ is downward crossing and therefore provides a trivial nil lower bound, which is satisfied by all rotors of finite mass. This bearing is therefore always stable for all rotors operating at $A = 10$. For $A = 20$, the intercepts respectively at $\Omega_{\text{neutral}}\omega = (0, 162.4\text{Hz})$ are upper and lower bounds for instability. The intercept at $|\Omega| \rightarrow \infty$ is again a trivial nil lower bound. Stable operation is limited to $m < 4.23\text{Kg}$.

The stability theorem has been stated for a one-dimensional problem, e.g., axial perturbation. The translational and angular *DOFs* involve coupling between $(\delta x, \delta y; \xi, \eta)$ according to (30) and (32). The equivalent

one-dimensional problem is readily deduced if the perturbation is at a concentric steady-state. The small eccentricity perturbation results have isotropic properties because of inherent rotational symmetry; e.g., for the translational *DOF*

$$\begin{aligned} Z_{xx}|_{\text{isotropic}} &= Z_{yy}|_{\text{isotropic}} = U \\ Z_{xy}|_{\text{isotropic}} &= -Z_{yx}|_{\text{isotropic}} = jV \end{aligned} \quad (39)$$

Equation (29) is reduced to

$$\begin{aligned} (m\omega^2)\Omega^2 &= U \pm jV \\ V(\Omega_{\text{neutral}}) &= 0 \\ (m_{\text{threshold}}\omega^2)\Omega_{\text{threshold}}^2 &= U(\Omega_{\text{neutral}})0 \end{aligned} \quad (40)$$

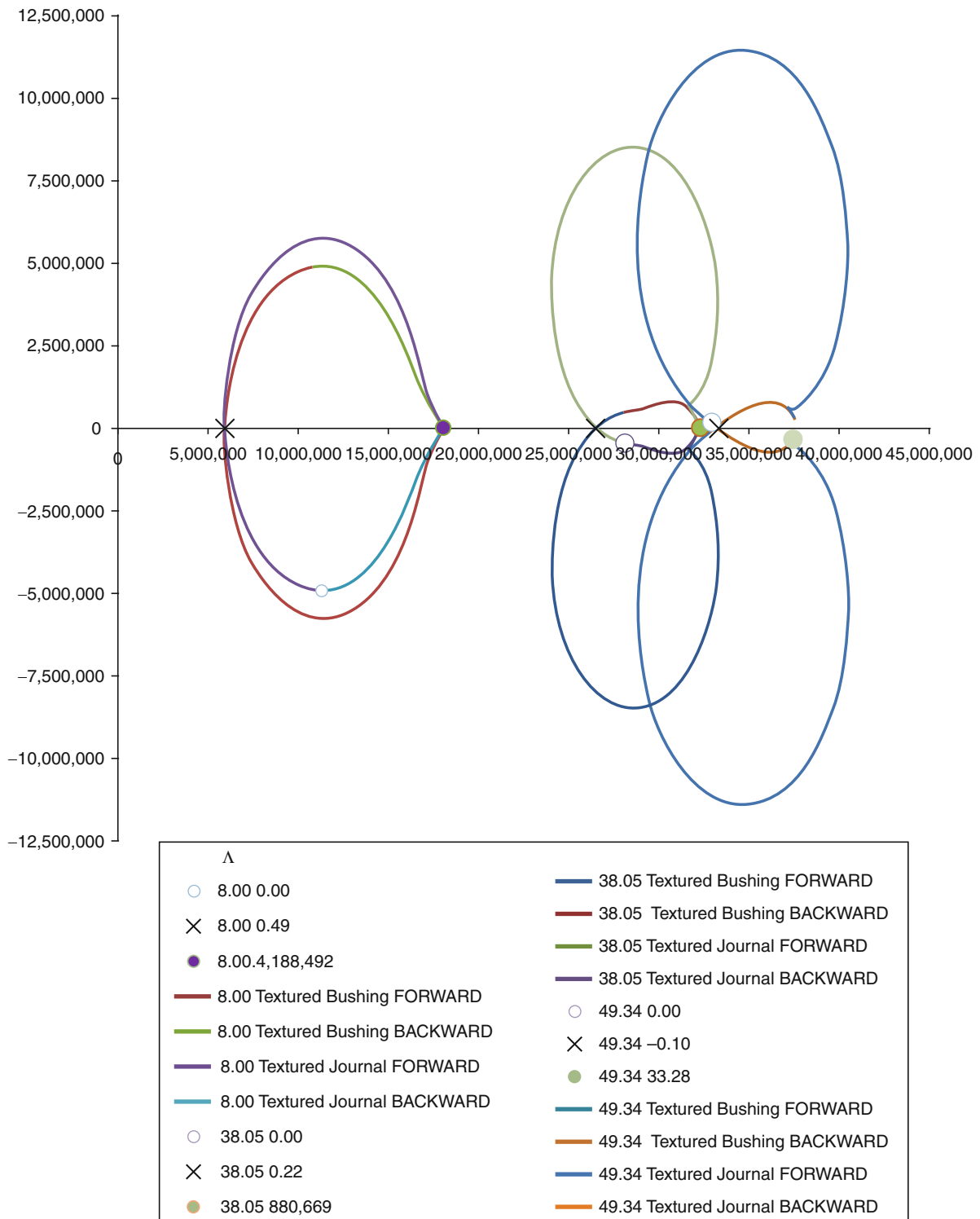
The alternative signs indicate the existence of two roots of the characteristic determinant that correspond to the alternative characteristic vectors that describe forward- and backward-rotating cylindrical whirl orbits, respectively (see Keating and Pan (1968) in list of references):

$$\delta y = \mp j \delta x \quad (41)$$

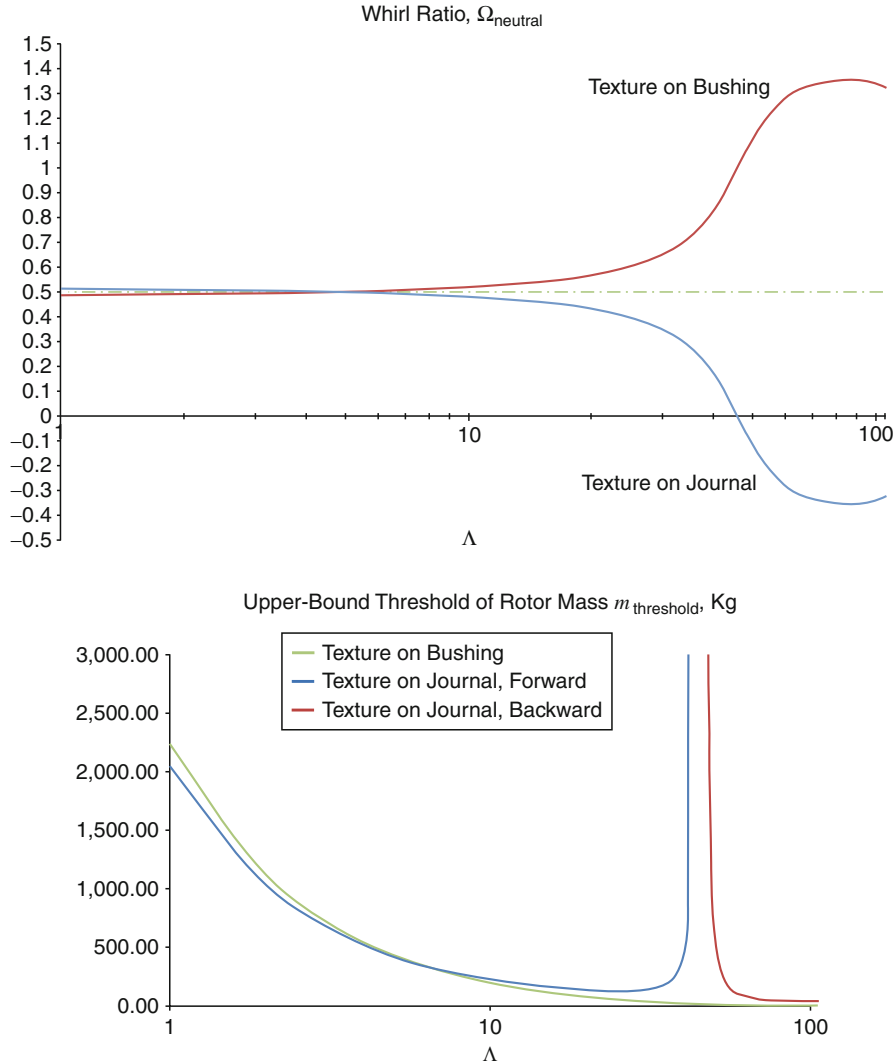
Distinction from the axial *DOF* comprises

1. The complex notation is to a spatial phase indicator. The perturbation film pressure is solved in a rotating coordinate system.
2. The alternative signs are associated with two wholly different sets of impedance solutions that may contain separate information regarding the state of stability.

Full-frequency compilation to search for all neutrally stable states requires computation of (U, V) with Ω stepped from a nil value to a suitably large value of both



Self-Excited Gas Bearing Instabilities, Fig. 7 Impedance contours of herringbone journal bearing



Self-Excited Gas Bearing Instabilities, Fig. 8 Stability parameters of herringbone journal bearing

algebraic signs. Completeness of the compilation is indicated by the asymptotic trend:

$$\lim_{|\Omega| \rightarrow \infty} U + jV = U_{\infty} + j\Omega^{-1} V_{\infty} \quad (42)$$

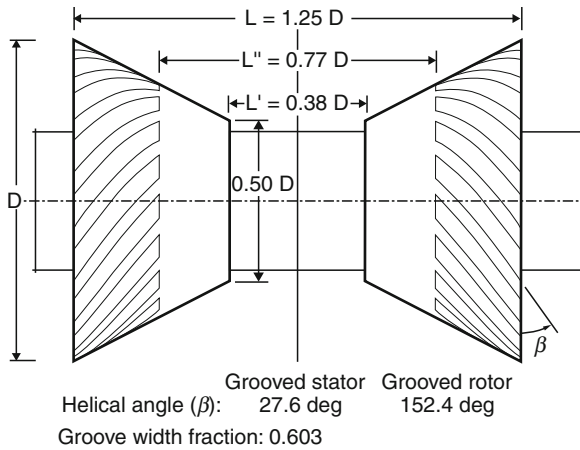
Stability characteristics of a herringbone journal bearing, Fig. 6, are reviewed here. In this design, pressure inducing textures are contained in 25% axial extent connected to the ambient at each end of the bearing. Gas viscosity is assumed to be that of atmospheric air at 50°C. Three values of Λ are selected for contour plotting to illustrate the unique features of this bearing as shown in Fig. 7. At each Λ there are two contour plots, which are mirror

images of each other with the same set of texture parameters as captioned for the case of textured bushing. The same centered pressurization can be achieved with texturing on either surface; groove inclination would be reversed, changed to $\beta = 155^\circ$ if the journal is textured. Values of Ω at mirrored points obey a “half-frequency reflection rule”:

$$\begin{aligned} U|_{\text{textured bushing}}(\Omega) &= U|_{\text{textured journal}}(1 - \Omega) \\ V|_{\text{textured bushing}}(\Omega) &= -V|_{\text{textured journal}}(1 - \Omega) \end{aligned} \quad (43)$$

The contours for the textured journal are marked with (\circ , \times and \bullet), respectively, for $\Omega = 0$, $\Omega|_{V=0}$ and

the largest of the data set. The largest $(2\pi)^{-1}|\Omega\omega|$ is 2×10^7 Hz for $\Lambda = 8.00$ and 38.05; contour closure is seen for these cases. $(2\pi)^{-1}|\Omega\omega|$ terminates at 9,800 Hz for

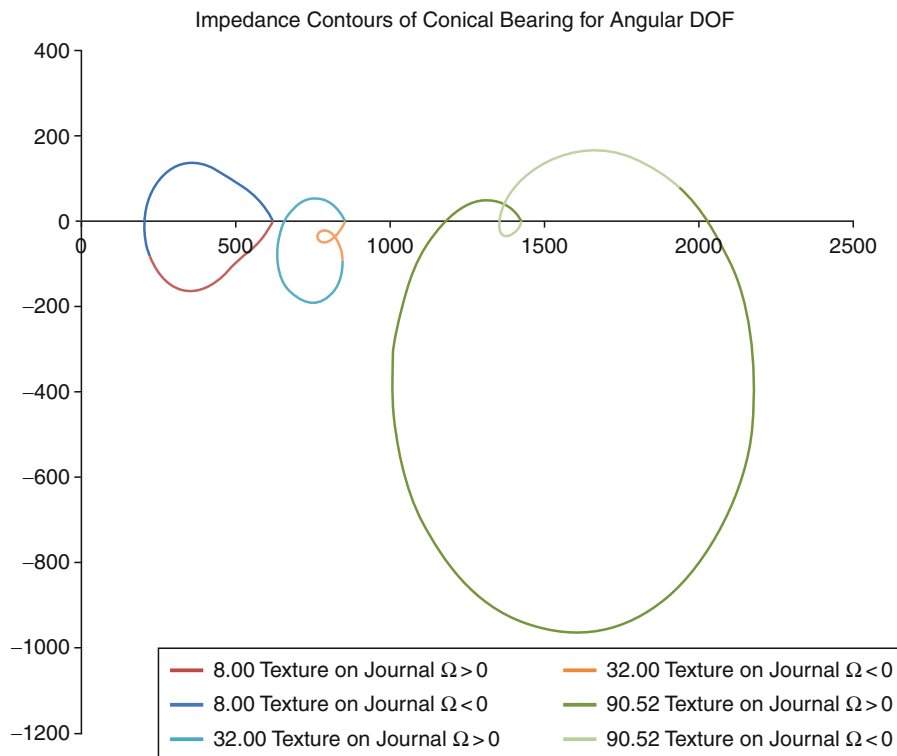


Self-Excited Gas Bearing Instabilities, Fig. 9 Inward-pumping conical bearing

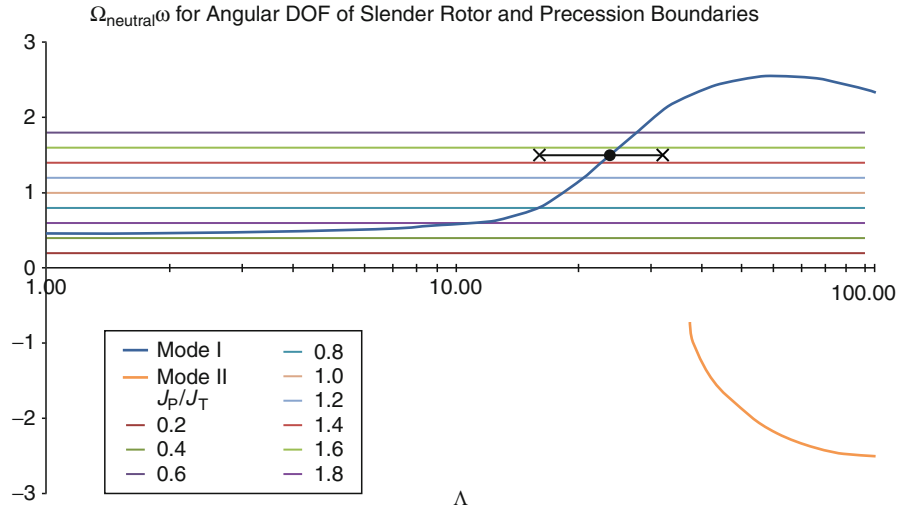
the data set of $\Lambda=49.39$; the contours do not close, although the asymptotic trend appears to be approaching and there isn't any non-trivial Ω_{neutral} . not accounted for. The stability theorem is applicable to all contours in Fig. 7. All intercepts on the real axis satisfy necessary and sufficient conditions required by (40).

Complete charting of stability characteristics of the herringbone journal bearing is shown in Fig. 8. Symmetrical placement of Ω_{neutral} on either side of 0.5 displays the "half-frequency reflection rule" stipulated by (43). The two curves stay close together until about $\Lambda \approx 20$. Thereafter, they diverge rapidly; especially as $\Omega_{\text{neutral}}|_{\text{texture on journal}}$ becomes negative beyond $\Lambda = 45.25$, there is considerable difference in the upper-bound $\mathcal{M}_{\text{threshold}}$ as shown in the lower pane for the bearing on one end. For instance, $\Lambda = 38.0$ corresponds to a rotational speed of 13,600 rpm; total upper-bound rotor mass of the shaft would be 536 kg with texture on the journal surface but is only 36 kg with texture on the bushing surface.

The small eccentricity perturbation problem of the angular DOF can be similarly reduced to an equivalent



Self-Excited Gas Bearing Instabilities, Fig. 10 A nomograph for determination of the neutrally stable state for the angular DOF



Self-Excited Gas Bearing Instabilities, Fig. 11 Impedance contours of inward-pumping conical bearing for angular *DOF*

one-dimensional system by making use of a pivoted rotating coordinate system illustrated in Fig. 3. In lieu of dealing with J_P directly, one can introduce the inertia ratio (J_P/J_T) to allow its inclusion during the process of searching of the neutral stability state; thus (31) is reduced to

$$V(\Omega_{\text{neutral}}) = 0, \quad [1 \mp \Omega_{\text{neutral}}^{-1}(J_P/J_T)] > 0,$$

$$U_{\text{neutral,equivalent}} = U(\Omega_{\text{neutral}}) \pm J_P \omega^2 \Omega_{\text{neutral}}$$

$$\text{then; } J_{T,\text{threshold}} = (\Omega_{\text{neutral}} \omega)^{-2} U_{\text{neutral,equivalent}} \quad (44)$$

Precession d'Alembert moment is combined with $U(\Omega_{\text{neutral}})$ to form $U_{\text{neutral,equivalent}}$ for computation of $J_{T,\text{threshold}}$.

Regardless how stability analysis is influenced by gyroscopic precession, it is necessary to find the intercepts of the impedance contour; shown in Fig. 9 is the profile of an inward-pumping conical bearing for which full-frequency survey of angular perturbation impedance has been compiled.

The inertia ratio of a rotor can be very small, the effect of gyroscopic precession would be minimal, or can approach 2.0 for a disk like structure, the effect of gyroscopic precession would be very pronounced. Its role in stability analysis for the angular *DOF* is two-fold: firstly, a negative value of $1 - \Omega^{-1}(J_P/J_T)$ would reject the possibility of a neutrally-stable state even if $V(\Omega)$ vanishes; secondly, for all identified neutrally

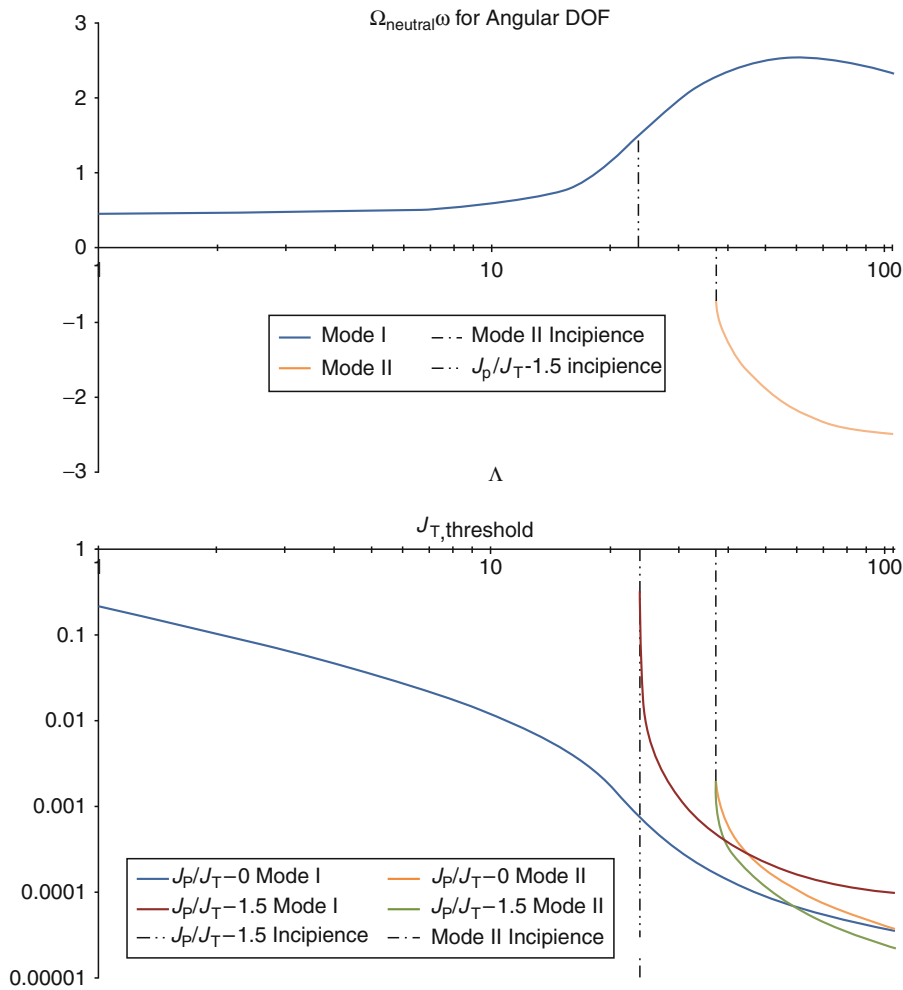
stable states, the threshold value of the inertial parameter is calculated as

$$J_{T,\text{threshold}} = (\Omega_{\text{neutral}} \omega)^{-2} U_{\text{neutral,equivalent}} \quad (45)$$

A nomograph procedure that combines the curves of slender rotor chart for Ω_{neutral} with lines of graduated J_P/J_T is a very simple to implement. An example of such a nomograph is shown in Fig. 10:

This particular bearing has two modes for vanishing $V(\Omega)$ at large λ ; they are revealed by the presence of four intercepts of the closed complex impedance contour in Fig. 11 for $\lambda = 90.52$. The incipience condition for Mode II is $\lambda = 37.34$. To examine Mode I regarding the effect of gyroscopic precession, horizontal graduation lines represent values of J_P/J_T ranging from 0.2 to 1.8 that should include the inertia property of common rotors. Given the value of J_P/J_T , all $\Omega|_{V=0} > 0$ below the graduation line are not accepted as Ω_{neutral} . For instance, if of interest is a rotor with $J_P/J_T = 1.5$, a short line is drawn to intercept $\Omega|_{V=0} > 0$, as marked between \times ; the intercept marked as \bullet is found at $\lambda = 23.76$. The final stability charts for the rotor with $J_P/J_T = 1.5$ are shown in Fig. 12. Incipience of Mode I $J_{T,\text{threshold}}$ is seen to be very abrupt.

If the steady-state eccentricity is finite, stability analysis for translational and angular *DOFs* would have to deal with (30) and (32). Cheng and Trumpler (1963) and Castelli and Elrod (1965) independently dealt with stability analysis of the infinitely long plain journal bearing under steady static load. Cheng and Trumpler used



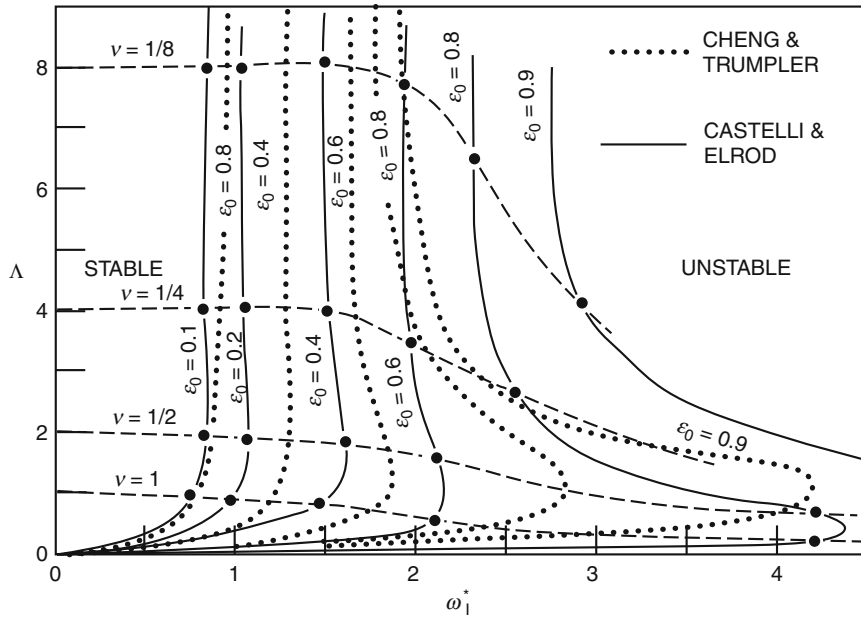
Self-Excited Gas Bearing Instabilities, Fig. 12 Stability charts for angular DOF with $J_p/J_T = 1.5$

Galerkin's method to approximate the film pressure profile with polynomials and succeeded to adapt Routh's criterion to determine the threshold speed of stability. Castelli and Elrod used finite difference approximation to compute the time-dependent film pressure equation; they combine a small perturbation analysis to find the neutrally stable state and a nonlinear orbit computation to determine whether the initial disturbance tends to grow or to decay. Essence of these earlier works is summarized in the stability map shown in Fig. 13.

Cheng and Pan (1965) extended the approach of Cheng and Trumpler by using two-dimensional profile functions to consider bearings of finite length. A typical set of stability map, pairing whirl frequency ratio with threshold speed for instability incipience, for $L/D = 1/4$ is reproduced here as Fig. 14. A prominent feature of the

threshold speed curves with fixed eccentricity is the presence of a peak at $\Lambda \approx 1.0$ for the infinitely long bearing and $\Lambda \approx 25.0$ for $L/D = 1/4$. Unfortunately, the latter observation is clouded by inability to carry out a more thorough study due to the inherent limitation of using truncated profile functions in Galerkin's method; dashed lines are estimated trends where computed results appear to be faulty.

It appears that availability of a robust computation method to deal with both the steady-state problem, (3), and the perturbation problem, (17), is a prerequisite to bringing to conclusion to this well-defined problem. Successful adaptation of Harrison's slider integrals to compute the air film slider by Diego (1987) suggests that a similar approach may be possible in stability analysis of gas film devices.



Self-Excited Gas Bearing Instabilities, Fig. 13 Stability map of an infinitely long plain journal bearing under eccentric steady state (After Castelli and Elrod 1965)

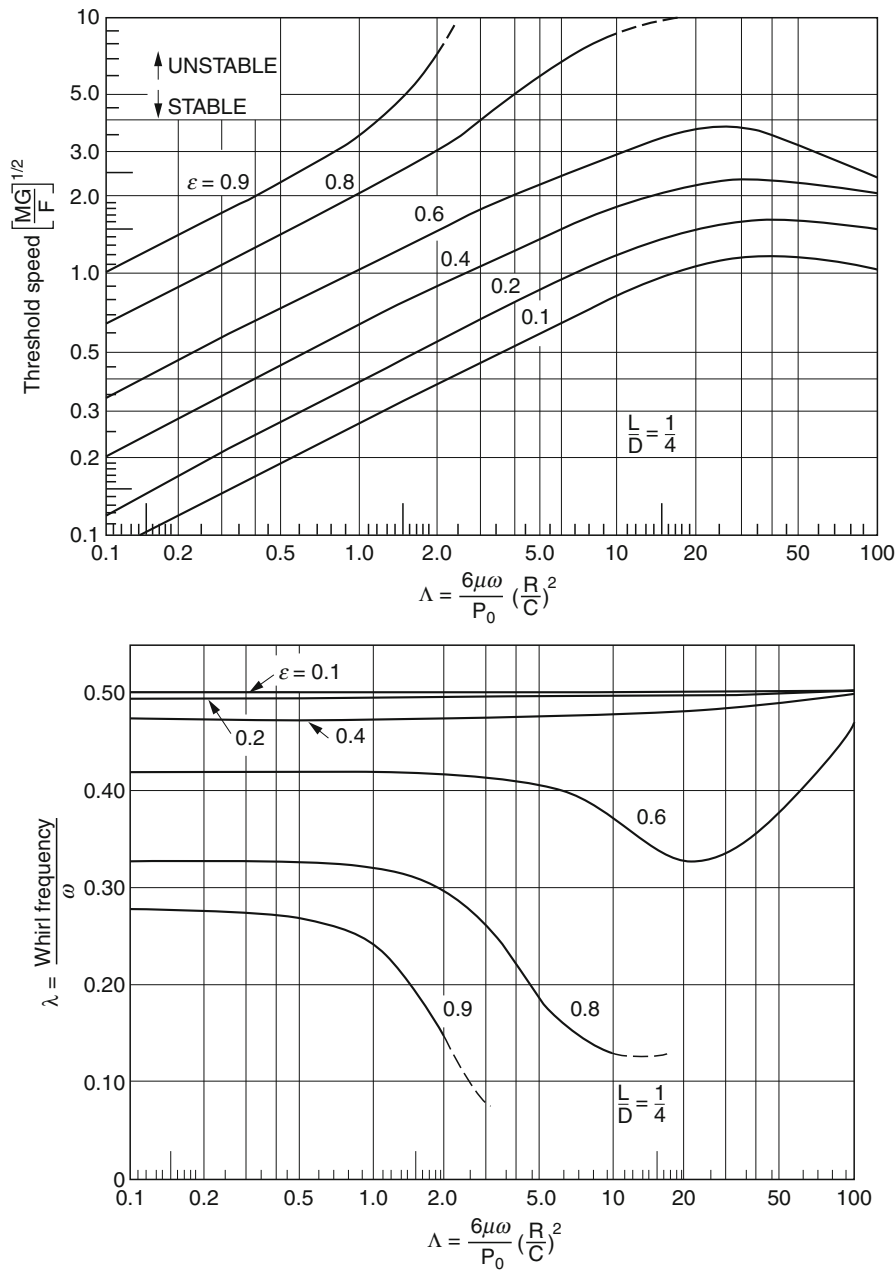
Key Applications

All gas bearings
MEMS
Design auditing
Trouble shooting
Comparison of alternative concepts
Evaluation of untested operating/environment

Nomenclature

Roman letters	
c	Damping coefficient, $N-s/m$
C	Radial clearance, m
D	Journal bearing diameter, m
δe	Total radial displacement, $= \sqrt{\delta x^2 + \delta y^2}$, m
F	Applied force, N
F_x, F_y, F_z	Cartesian components of steady-state bearing force, N
h	Film thickness, m
H	Normalized film thickness, $= h/C$
H_0	Steady state H
k	Stiffness coefficient, N/m
$\bar{i}, \bar{j}, \bar{k}$	s of ground-fixed Cartesian coordinate system
J_p	Polar moment of inertia of rotor, $kg - m^2$

J_T	Transverse moment of inertia of rotor, $kg - m^2$
L	Axial length of journal bearing, m
m	Rotor mass, kg
M_ξ, M_η	Components of moment vector about respective ground-fixed transverse s
p	Film pressure, MPa
p_a	Ambient pressure, MPa
p_h	Film perturbed pressure coefficient, (8), MPa/m
r_i, r_o	Inner-, outer-radii of thrust bearing, m
R	Journal bearing radius, $= \frac{1}{2}D$, m
dS	Differential area element of bearing surface, m^2
s	Complex Laplace exponential coefficient, $\equiv \lambda + j\Omega\omega$, s^{-1}
t	Time, s
U_{xx}, V_{xx} etc.	Real and imaginary parts of Z_{xx} etc., N/m
\bar{V}	Entrainment velocity, m/s
$\Delta \bar{V}$	Velocity difference between upper and lower surfaces, m/s
x_0	Static equilibrium position, m
\dot{x}, \ddot{x}	First and second time derivatives of x , m/s^2
\bar{x}	Amplitude of displacement with exponential time-dependence, m



Self-Excited Gas Bearing Instabilities, Fig. 14 Stability map of plain journal bearing under eccentric steady state, $L/D = 1/4$ (After Cheng and Pan 1965)

$\delta x, \delta y, \delta z$	Infinitesimal linear displacement components of rotor, m
Z_{xx}, Z_{yx} etc.	Radial perturbation bearing impedance, (11), N/m
Z_{zz}	Axial perturbation bearing impedance, $\equiv -\delta F_z / \delta z = U_z + jV_z, N/m$

$U_{z,\infty}, \Omega V_{z,\infty}$	Real and imaginary part of $\lim_{ \Omega \rightarrow \infty} \delta F_z / \delta z$
Greek Letters	
α	Width fraction of grooves of textured bearings
β	Inclination angle of texture pattern, $^\circ$
Γ	Ratio of groove depth to bearing gap

$\vec{\phi}$	Volumetric film flux vector, m^2/s
$\delta\xi, \delta\eta$	Infinitesimal angular displacement components of rotor, <i>radians</i>
λ	Real part of s
A	Bearing number, $= 6(\mu VR)(p_a C^2)^{-1}$
μ	Gas viscosity, $Pa \cdot s$
\bar{v}	Outward normal base-vector of bearing surface
$\sigma_\tau, \sigma_C, \sigma_P$	Rarefaction coefficients for mid-film shear, Couette flux, Poiseuille flux
$\bar{\tau}$	Mid-film in-plane shear stress, <i>MPa</i>
ω	Rotational rate, <i>rad/s</i>
Ω	Ratio of oscillation rate to rotational rate, <i>whirl ratio</i>
Ω'	Damped natural frequency, <i>rad/s</i>
Mathematical symbols	
div	Divergence operator, $\equiv \text{grad} \cdot, m^{-1}$
grad	Gradient operator, m^{-1}
$h_{()}$	Jacobian derivative of film thickness with respect to the generic displacement ().
$\delta()$	Infinitesimal deviation of ()
j	Identifier of the imaginary part of a complex number, $= \sqrt{-1}$
$\bar{\nabla}$	Normalized gradient operator
Subscripts	
0	Pertaining to the steady-state
neutral	Pertaining to the neutral state of stability
m	Pertaining to the radius of the circle that separates textured and smooth annuli of spiral grooved thrust bearing, Fig. 4
x, y, z	Pertaining to linear perturbation along the respective Cartesian directions
ξ, η	Pertaining to angular perturbation about respective ground-fixed transverse s

References

A. Burgdorfer, The influence of the molecular mean free path on the performance of hydrodynamic gas-lubricated bearings. *J. Basic Eng.* **80**(1), 94–100 (1959)

V. Castelli, H.G. Elrod, Solution of the stability problem for 360 degree self-acting, gas lubricated bearings. *J. Basic Eng.* **87**(1), 199–212 (1965)

H.S. Cheng, C.H.T. Pan, Stability analysis of gas-lubricated, self-acting, plain, cylindrical, journal bearings of finite length, using Galerkin's method. *J. Basic Eng.* **87**(1), 185–192 (1965)

H.S. Cheng, P.R. Trumpler, Stability of the high-speed journal bearing under steady load. *J. Eng. Ind.* **8**(5), 274 (1963)

J.A. Diego, An unconditionally stable, high resolution algorithm for gas lubricated problems, Doctoral Dissertation, Columbia University, UMI Order Number 8724042, 1987

H.G. Elrod, Jr., A. Burgdorfer, Refinement of the theory of gas-lubricated journal bearing of infinite length, in *Proceedings, First International Symposium on Gas-Lubricated Bearings*, October 1959, U. S. Government Printing Office, ACR-49 (Office of Naval Research, Washington DC, 1959), pp. 93–118

W.J. Harrison, The hydrodynamic theory of lubrication with special reference to air as a lubricant. *Trans. Cambr. Phil. Soc* **XXii**, 6–54 (1913)

S.-C. Kang, A kinetic theory description for molecular lubrication, Ph.D. Dissertation, Carnegie Mellon University, UMI Number 9801040, 1997

W.H. Keating, C.H.T. Pan, Design studies of an opposed-sphere gyro spin-axis gas bearing. *J. Lubr. Technol.* **90**(4), 753–760 (1968)

W.-L. Li, A database for Couette flow rate considering the effects of non-symmetric molecular interactions. *J. Tribol.* **124**(4), 869–873 (2002)

S.B. Malanoski, C.H.T. Pan, “The Static and Dynamic Characteristics of the Spiral-Groove Thrust Bearing”, *J. Basic Eng.* **87**(3), 547–558 (1965)

C.H.T. Pan, Spectral analysis of gas bearing systems for stability studies, in *Developments in Mechanics*, eds. by T.C. Huang and M.W. Johnson, Jr. Dynamics ad Fluid Mechanics, Proceedings of the Ninth Midwestern Mechanics Conference, Madison, vol. 3, Part 2 (Wiley, New York, 1965), pp. 431–448

Self-Excited Shaft Whirl

► [Self-Excited Gas Bearing Instabilities](#)

Self-Lubricating Bearings

► [Porous Metal Journal Bearings](#)

Self-Lubricating Coatings and Composite Coatings

► [Self-Lubricating Treatment of Light Alloys](#)

Self-Lubricating Hard/Ultra-Hard Coatings

G. A. ZHANG¹, LIPING WANG²

¹Lanzhou Institute of Chemical Physics, Lanzhou, People's Republic of China

²State Key Laboratory of Solid Lubrication, Lanzhou Institute of Chemical Physics, Chinese Academy of Science, Lanzhou, People's Republic of China

Synonyms

[Hard self-lubricating coatings](#); [Hard/ultra-hard self-lubricating coatings](#)

Definition

Self-lubricating hard/ultra-hard coatings is a class of coatings, which usually exhibit low friction coefficients and relatively high hardness. The threshold of the hardness is roughly estimated to be around 20 GPa. And the threshold value of friction coefficient about 0.3 between solid lubricants and anti-wear coatings, close to typical friction values of metals.

Scientific Fundamentals

Introduction

Increasing demands for higher power density, longer durability, and greater efficiency in future mechanical systems are pushing current materials and coatings to their limits. Modern mechanical assemblies often operate under boundary-lubricated sliding regimes where asperity interactions can take place, as opposed to a thick fluid film separating the sliding surfaces and thus preventing direct contact. Furthermore, some industrial applications even require sliding contact without any type of liquid lubrication, thus exacerbating the situation. To achieve and maintain higher efficiency and durability under such increasingly more severe sliding conditions, protective and/or solid lubricant coatings are becoming prevalent. Low friction coefficients can effectively reduce the contact temperatures during boundary-lubricated and/or dry sliding and thus have a huge potential to reduce the thermal load for coating and substrate material. A lot of effort is being made to develop low-friction protective layers to act as solid lubricants with the goal to replace the commonly used expensive and hazardous coolant lubricants. Especially in high speed and dry cutting applications, low-friction and lubricating mechanisms of the coating itself are required in addition to excellent mechanical properties. However, the present commercially used hard transition metal nitride or carbide coatings lack lubricating properties and show, consequently, relatively high friction coefficients in the range between 0.4 and 1.0 against steel. Reducing the friction coefficient while retaining the high hardness and wear resistance of a hard coating, which means self-lubricating hard/ultra-hard coatings, could be achieved by the development of new coating concepts, architectures, and deposition methods in the field of hard coatings technologies (Mayrhofer et al. 2006).

The coating with immense potential in mechanical application is diamond-like carbon (DLC), which has both high hardness (often exceeding 20 GPa) and low friction coefficient. However, the strong dependence of the tribological properties of DLC on the environment

(especially on the presence of oxygen and water vapor), the significant stress levels in the coatings, the high wear rates when sliding against ferrous materials and degrading in high temperature (below 350°C) are the among the most serious disadvantages of the DLC coatings. Intrinsic solid lubricants like metal oxides (PbO, TiO, NiO, CoMoO₄), inorganic fluorites (CaF₂, BaF₂), lambda structure material (transition metal dichalcogenide, MoS₂, WS₂, and graphite), or soft metals (Ag, Au, In) have lower hardness and poor wear resistance, which does not make them ideal for surface engineering applications in terrestrial environment. Currently, there is a challenge to acquire self-lubricating hard coatings with both high hardness coatings and low friction coefficient. Advanced coating approaches combine multiple materials in a single composite coating in an attempt to circumvent the shortcomings of individual solid lubricant materials (Hogmark et al. 2000).

The Classification of Self-lubricating Hard Coatings

Carbon-Based Self-lubricating Hard Coatings

Advances in coating science and technology have led to the introduction of hard and lubricious carbon-based coatings, such as diamond-like carbon (DLC), carbon nitride (CN_x), silicon carbide, boron carbide (nominally B₄C), and nanostructured DLC-based nanocomposites as preferred candidates for self-lubricating hard coatings (Robertson 2002). However, the friction and wear properties of these carbon-based coatings depend strongly on the environment. For example, during dry sliding between two hydrogenated DLC-coated surfaces in nominally dry nitrogen ambient, the friction coefficient is sensitive to the presence of water vapor and increases with time delay between sequential sliding experiments. On the other hand, in the presence of oxygen, acceleration of graphitization (i.e., conversion of sp³- to sp²-bonded carbon atoms) and oxidation of carbon may occur, leading to increased friction and wear. In some instances, counterface materials react with the carbon coating, forming softer components and resulting in wear of a nominally harder surface by a softer counterface. The environment and surface chemistry determine the friction and wear performance of carbon-based coatings.

Diamond-like carbon (DLC) is an amorphous carbon (a-C) or hydrogenated amorphous carbon (a-C:H) thin film material with a high fraction of sp³ carbon bonding. The great variety of DLC structures and compositions leads to a wide range of mechanical properties and

performance. The hardness varies from a few GPa up to more than 60 GPa, while the elastic modulus ranges from several tens of GPa up to several hundreds of GPa. Friction coefficients of DLC films, which typically range from 0.01 to more than 0.5, depend on the nature of the film and the conditions used for friction testing, such as test conditions (load and speed), test environment, temperature, and counterface material. And the friction behavior of DLC is controlled by an interfacial transfer layer formed during sliding. This transfer layer of low shear strength (sp^2 -type) is formed from the surface of the DLC coating and is responsible for the low friction coefficients. In ambient humid air at a relative humidity of 20–60%, hydrogen-free DLC films generally exhibit lower friction coefficients (<0.15) compared with hydrogenated DLC film. In a-C films, friction easily causes a local shear-induced graphitization (sp^3 – sp^2) in the contact zone as the high contact flash temperatures, resulting in reduced friction due to the formation of a thin graphitized tribolayer. The increase of the friction coefficient with increasing humidity can be explained by a condensed water layer at asperity contacts, which has a “cooling effect” at the contact zone, so that necessary temperatures may not be attained and, therefore, the graphitization process is expected to be suppressed. At very high humidity, no graphitization formation results in relatively high friction coefficients. In inert environments such as dry nitrogen and vacuum, the amount of hydrogen in the coating structure determines the tribological performance of the DLC coating, where the hydrogen atoms determine the contact bonding between the DLC film and the counterface. For hydrogenated DLC, friction coefficient depends strongly on the relative humidity. The friction coefficient below 0.05 is found in a vacuum and at low humidity and increases strongly at high humidity. Contact with a different surface causes a transfer layer of hydrogenated DLC to be formed on the other surface. Thus, the contact is between two basically similar hydrophobic hydrogenated DLC surfaces, and the friction coefficient is very low. High humidity is believed to interfere with the formation of the contact layer and cause it to become oxidized/hydrated, so that it no longer forms the van der Waals bonds. If the transfer layer does not form, the counter surface is no longer hydrophobic, and the friction coefficient is much higher (Donnet and Erdemir 2008).

To overcome the disadvantageous humidity sensitivity of DLC, metal atoms in the range of 10–40% can be added to the DLC film to obtain less humidity degradation compared with pure DLC films. DLC films have been made with many different metals, including Ti, Nb, Ta, Cr, Mo, W, Ru, Fe, Co, Ni, Al, Cu, Au, and Ag, mainly by

sputtering of metallic targets in the presence of acetylene or other hydrocarbon gas (Donnet and Erdemir 2008). In many cases, metals are in the form of small nanocrystallites of pure metal or metal carbide (depending on the nature and concentration of the metal) dispersed throughout the carbon network. Interesting optimum tribological properties may be obtained for each type of metal dopant, with a concentration that seems to depend on its nature. However, the tribological performances of the Me-doped DLC coatings can differ sharply from one another. For instance, addition of metals like W, Ti, Cr, Ta to DLC films forming nanocrystalline metal carbides embedded in the DLC matrix thus improve their performance in humid and dry conditions. Tribological tests in ambient air conditions always exhibit steady-state friction in the range 0.10–0.25 with a slight dependence on humidity and load for metal contents below 30 at.%. The addition of silicon to a-C:H improves its friction properties by allowing it to maintain its low friction coefficient from low up to high humidity. It is believed that this occurs because the Si changes the nature of the transfer layer, forming a silica-gel-like sacrificial layer. Furthermore, the real response of a coating in a specific application has to be determined by conducting field tests. To further overcome these variations in performance and to extend the applicability over a broader range of test conditions, researchers have been exploring novel coating architectures having multilayers, nanostructures, or composites. This can cause a coating that is optimized for a certain tribological applications to perform worse under other tribological conditions.

Amorphous carbon nitride (CN_x) is suggested as a superhard solid lubricant as the interesting mechanical properties of α -C₃N₄. The hardness of these coatings typically varies between 15 and 30 GPa. The friction coefficient of the steel-CN_x pair in unlubricated conditions was about 0.16. Similar to DLC, friction and wear properties of CN_x coatings are dependent on the atmospheric conditions, degrading with increased humidity in the testing environment. In pure nitrogen (without oxygen and water vapor), friction coefficients <0.01 are obtained with smooth surfaces and when both surfaces are coated with amorphous CN_x, either initially or through the formation of transfer films. Amorphous silicon carbide synthesized by plasma methods has relatively high hardness range from 10 to 25 GPa. The lubricating performance of this amorphous silicon carbide film is greatly improved by annealing at 800°C. When annealed at 800°C, all remaining hydrogen was liberated from the film. The dehydrogenation process results in the formation of graphite-like structures, leading to lower friction.

The low friction (friction coefficient ~ 0.01) was maintained until the film was worn out. Boron carbide (B_4C) films are characterized by high hardness (ranging from 25 to 45 GPa) and stress resulting in excellent wear resistance, however, they always show relatively high friction coefficients of 0.3–0.4. After annealing at 800°C, reduced friction coefficients of 0.03–0.05 due to oxidation of B_4C . This low-friction mechanism is based on the reaction of the boric oxide (B_2O_3) with ambient humidity to form a thin boric acid (H_3BO_4) film. The low friction coefficient of boric acid is associated with its layered triclinic crystal structure. The layers consist of closely packed and strongly bonded boron, oxygen, and hydrogen atoms, but the layers are widely separated and attracted by van der Waals forces. During sliding, these atomic layers can align themselves parallel to the direction of relative motion and slide easily over one another. The temperature sensitivity of boric acid, however, restricts the applications of low-friction B_4C coatings at elevated temperatures. Above 170°C, boric acid tends to decompose, thus losing its layered crystal structure and hence its lubricity. The use of boric acid as a low-friction coating is also limited to reasonably humid conditions, precluding its use in dry or vacuum applications (Erdemir et al. 1991).

Self-lubrication by Oxide Formation

Intrinsic solid lubricants like DLC, MoS_2 , and h-BN often begin to fail in their tribological effectiveness with increasing temperature, in humid atmosphere, or due to oxidation. However, several approaches have been suggested to improve the friction properties of hard coatings by self-adaptive lubrication mechanisms occurring during application, e.g., during dry cutting or by tailoring their oxidation behavior. Conventional lubricants that are a product of chemical reactions between coating and moisture of the ambient atmosphere, e.g., the formation of boric acid in B_4C , always begin to fail in their tribological effectiveness with increasing temperature, limiting applicability at elevated temperature (Erdemir et al. 1991).

Thus, a new concept of high-temperature lubrication was found in the use of lubricious oxide materials with easily moveable shear planes, also referred to as Magnéli phases. These phases exhibit good thermal stability, high resistance against tribo-oxidation, and low adhesion. The oxides of W, Mo, V, and Ti especially form homologous series with planar faults according to the common Magnéli phase principles Me_nO_{2n-1} , Me_nO_{3n-1} , and Me_nO_{3n-2} (Woydt et al. 1998; Mayrhofer et al. 2006). Such crystalline structures based on the rutile structure contain rutile-like chains of edge-sharing octahedral, interrupted by shear planes every n th octahedron.

Generally, these shear planes exhibit reduced binding strength. In comparison to common solid lubricants such as MoS_2 , where every second layer offers crystallographic slip ability, Magnéli phases might exhibit less promising lubrication performance, because here only every n th layer has a crystallographic shear structure. Solid lubricants based on these Magnéli phase oxides have not yet found wide use because of the difficulty in achieving and maintaining the very narrow range of oxide stoichiometry, which is necessary for good lubricity.

As VN is a potential candidate to add self-lubricious properties to existing hard thin films for advanced cutting applications, some effort has been made to develop TiAlN/VN superlattice films where the beneficial properties of both types of layers are combined (Mayrhofer et al. 2004). TiAlN/VN has proven to be an excellent candidate in protecting machine wear parts and cutting tools due to its high wear resistance and low friction. The mechanism of the low friction was laid on the formation and stability of V_2O_5 as it is known to transform to VO_2 via different V-oxides, which partly count as Magnéli phases. The TiAlN/VN superlattice films also achieve low friction coefficient at room temperature. During sliding, a V_2O_5 containing oxide with lubricious properties is formed at asperity contacts. The excellent tribological properties of TiAlN/VN superlattice films can thus be attributed to the formation of these particular oxides.

TiN deposited by PVD as well as by CVD exhibits a friction coefficient in the range of 0.4–0.8 against steel counterparts. Using PACVD, the friction coefficient of TiN could be decreased to a value of 0.17. These low values are attributed to a certain amount of chlorine in the coatings, which is incorporated as $TiCl_4$ and is used as a chloridic precursor in the PACVD process. Moreover, chlorine implantation into TiN coatings also decreases the wear loss and the friction coefficient. For coatings with a chlorine content below 3.2 at.%, the friction coefficient is stable at a value of 0.75 over the test period, whereas for higher chlorine concentrations a sharp drop to a value of 0.17 occurs. Low chlorine impurities are assumed to be incorporated into the TiN phase, whereas for concentrations exceeding 3.2 at.% chlorine also segregates to grain boundaries. PACVD TiN coatings exhibit grain refinement by continuous renucleation during growth, whereas, on the other hand, the formation of a thin rutile layer, in humid air, is stimulated at the topmost surface (Badisch et al. 2004). This Cl-induced low-friction effect has been demonstrated for several transition metal nitride coating systems, e.g., TiN and nanocomposite Ti(B)N, where the Cl addition was provided by the PACVD process or by Cl ion implantation, respectively. Summing up,

the self-lubrication properties of Cl containing PACVD TiN coatings can be explained by the in situ formation of the easily shearable rutile phase in the contact zone between the sliding partners (Aizawa et al. 2004).

State-of-the-Art Structure Self-lubricating Hard Coatings

The Architectures of Self-lubricating Hard Coatings

The development of the technique to make a hard, self-lubricated coating is interesting. In order to improve self-lubricating behavior, different methods have been employed, building on the composite structures that involve the combination of soft lubricating phase(s) within hard, wear-resistant phase(s). The state-of-the-art structure designs of self-lubricating hard coatings have been addressed using three different approaches as follows: (a) inclusion of metal doping in solid lubricant coating, (b) solid lubricant coating layer on top of hard coating layer in the form of multilayer, and (c) inclusion of solid lubricant in a hard matrix (Hogmark et al. 2000). Figure 1 shows possible architectures of state-of-the-art structure designs of self-lubricating hard coatings.

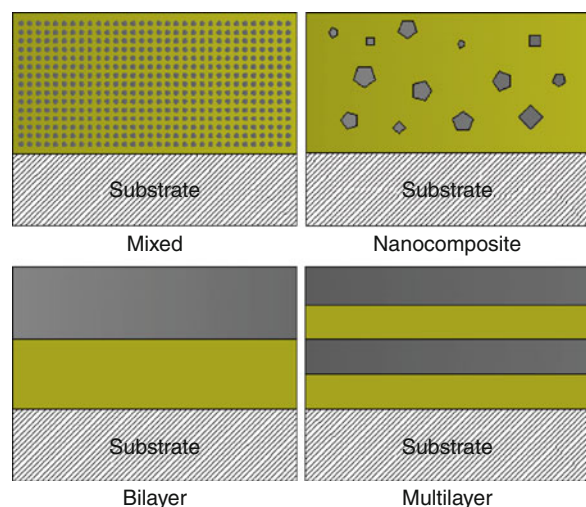
Mixed and Nanocomposite Coating Architectures

In the mixed coatings, the matrix and the dopant are mixed on an atomic level without any phase separation. The resulting solid solution is monophase and can be amorphous or crystalline. In the composite coatings at least two phases can be distinguished. Clusters (mono-

polycrystalline or amorphous) of the solid lubricant with several nanometers magnitude in size are uniformly dispersed within the hard phase matrix (amorphous or crystalline). Some solubility of the solid lubricant in the hard phase and vice versa is also possible. The solid lubricant is to be present throughout the entire thickness of coating in order to maintain the low friction coefficient during its entire period of service. Furthermore, the mechanism of solid lubrication in the mixed and nanocomposite coatings can be easily predicted. As in the nanocomposite coating wear process, there are always reservoirs of solid lubricant in the nanocomposite coatings exposed at the surface of the wear track. These reservoirs provide solid lubricant, thus decreasing the friction coefficient. For the mixed coatings, when the coatings are rubbed, some re-orientation and re-crystallization of the coating materials takes place so that the coating surface might consist of a more crystalline material with a basal plane parallel orientation. Finally, low friction is achieved in the wear process.

Bilayer and Multilayer Coating Architectures

The bilayer and multilayer architectures are realized by the sequential of the constituents. The thickness of each individual layer can be controlled, from several nanometers to several micrometers. For the bilayer coating structure, once the top layer of the solid lubricant is worn, no further lubrication will be provided for the underlying hard phase. The multilayer structure can be useful in a real friction situation since such an effect, if constantly repeated many times throughout the multilayer coating wear life, can lead to the overall effect observed with the composite and probably the mixed coatings. If the individual layers of the solid lubricant and the hard phase are very thin, in the range of few nanometers, overlapping of the effects of each phase on the overall properties of the coating will take place. Furthermore, due to local fluctuations in friction conditions, the multilayer coatings never wear layer by layer in such a way that there is only one phase present at the entire surface of the wear track at any moment.



Self-Lubricating Hard/Ultra-Hard Coatings, Fig. 1 The architectures of self-lubricating hard coatings

The Classification of State-of-the-Art Structure Self-lubricating Hard Coatings

As mentioned above, the self-lubricating hard coatings consist of two components: a hard phase and a solid lubricant. The ideal candidate for the role of the hard phase in the self-lubricating hard coatings would possess good mechanical properties such as high hardness and low wear rate. Hard coatings, such TiN, TiC, Ti(C,N), (Ti,Al)N, (Ti,Al)(C,N), ZrN, CrN, Al₂O₃, diamond-like carbon (DLC) and others, possess high hardness (above 20 GPa),

low wear rate, and relatively high friction coefficients. Solid lubrication can be achieved using metal oxides (PbO, TiO, NiO, CoMoO₄), inorganic fluorites (CaF₂, BaF₂), lambda structure material (transition metal dichalcogenide, MoS₂, WS₂ and graphite), or soft metals (Ag, Au, In, Cu). This group of coatings has low friction coefficient and their wear-diminishing properties are brought about by providing solid lubrication. Significant disadvantages of the coatings in this group, however, are their relatively low hardness and low wear rate compared to the hard coatings. Classifying of self-lubricating hard composite coatings by lubrication phase is summarized below:

In the class of transition metal dichalcogenides, molybdenum disulphide (MoS₂) is the most popular lamellar solid lubricant. MoS₂ shows good friction performance in vacuum and under dry running conditions, but degrades quickly in moist and oxidizing environments. Friction coefficients of 0.002–0.05 can be observed in vacuum, dry, or inert atmospheres, which increase rapidly to 0.2 in humid air. MoS₂ coatings are soft (the hardness is about 4 GPa) with poor adhesion to metal substrate and, of course, are totally unsuitable for use in terrestrial conditions.

The development of the technique to make a MoS₂-based, hard, solid-lubricated coating is interesting. One way to achieve this goal is to deposit a soft-lubricated MoS₂ film on the hard films. MoS₂-based coatings usually deposit as a top layer on hard underlayers (such as TiN, TiC, CrN). A MoS₂-based coating on the hard CrN layer leads to a significant improvement in tribological properties. A maximum wear volume and coefficient of friction (COF) reduction of 95% is observed. However, when the top layer of the solid lubricant is worn, no further lubrication will be provided for the underlying hard phase. The decreasing friction and wear only act on the initial process when there is a layer of a solid lubricant on a hard coating. Since the bilayer coating system is practically improbable to obtain super low friction, hardness, oxidation, and wear resistance properties, another way to achieve this goal is to deposit a nanocomposite structure with the incorporation of a lubricious phase like MoS₂ into a harder matrix (e.g., TiN, TiB₂, CrN, CrB₂) in order to maintain a reservoir of solid lubricant throughout the coating thickness. Such hard solid lubricant film is effective in maintaining a coefficient of friction less than 0.3 through a long, stable lifetime without showing significant deterioration in hardness. The TiN–MoS_x composite coatings show very low friction characteristics (with friction coefficient less than 0.05) under normal atmospheric conditions. Incorporation of 8% MoS₂ in the TiN–MoS_x composite matrix

also shows significant reduction in friction coefficient (~ 0.2) while maintaining high hardness around 30 GPa. CrB₂–MoS₂ composite coating with 19% concentration of MoS₂ is attributed to the formation of an optimal nanocomposite structure and mechanical properties. In every case examined, separate MoS₂ peak is not found in the XRD spectra and hence it is postulated that the presence of MoS₂ is in the form of elemental Mo and S at grain boundary, which formed MoS₂ when put into a tribological application. The re-orientation and recrystallization of the elemental Mo and S with a more crystalline material with a basal plane parallel orientation in the wear surface cause the self-lubrication performance (Hogmark et al. 2000). Another effective method of increasing friction and wear properties of MoS₂ coatings is the addition of titanium atoms (up to ~ 20 at.%) to the MoS₂ structure. In this so-called MoSTTM structure (Teer 2001), titanium is thought to be in solid solution within the MoS₂. HRTEM investigations show that the Ti atoms actually also form nanocrystallites in a matrix of MoS₂ in the form of bundles of curved MoS₂ basal planes. The distortion due to the titanium atoms is responsible for an increase in hardness from ~ 4 GPa (pure MoS₂) up to 10 ~ 20 GPa, yielding improved wear resistance. Another advantage of titanium-doped MoS₂ compared with conventional MoS₂ is a reduced sensitivity of the coating to water vapor. Therefore, the applicability of these coatings is extended from dry conditions up to 50% humidity.

To combine the superior friction properties of DLC in humid air and WS₂ in dry conditions with the high wear resistance of transition metal carbides, a nanocomposite concept for tribological applications has been proposed by Voevodin et al. The nanocomposite coating within the W-C-S system consist of 1–2 nm WC and 5–10 nm WS₂ grains embedded in a DLC matrix. The WC/DLC/WS₂ nanocomposite exhibited chameleon-like self-adaptation to operations that occur in aerospace systems, providing friction reduction in both dry and humid environments (Voevodin and Zabinski 2005).

Soft metal coatings like lead, silver, gold, copper, nickel, and indium exhibit low shear strength over a large temperature range and excellent friction properties when kept sufficiently thin, but are susceptible to high wear, plastic deformation, and gross plowing during sliding causing rapid coating loss, increased friction, and irregular surface topographies. Inclusion of soft metals as solid lubricating phases in carbide, oxide and nitride matrix aims to improve tribological performance, as they possess sufficiently low shear strength as well as strong adhesion to the substrate, high film thickness, low surface roughness, enduring high load, and sliding speed.

The benefit of these nanocomposites, such as TiN/Ag, TiC/Ag, CrN/Ag, TiN/Cu, and yttrium-stabilized zirconia YSZ/Au systems, is that soft metals may act as lubricants at room temperature as well as high temperatures due to their low shear strength and stable thermochemistry.

Transition metal nitrides combined with soft metals are of particular interest as they are relatively easy to co-deposit by reactive magnetron sputtering and form nanocomposite structures, due to the lack of miscibility between the matrix and the lubricant. For instance, the incorporation of silver into the CrN coating tends to form a structure that consists of a matrix of CrN surrounding nanoparticles of silver, and offers the potential to modify the tribological properties. The self-lubricating nature of CrN/Ag coatings, combined with their high hardness and scratch resistance, makes them attractive for tribological applications (Mulligan and Gall 2005). Alloying TiN and TiC coatings with Ag also offers a high potential to enhance the tribological behavior in a wide temperature range and under varying environmental conditions. Ag containing nanocomposites is promising as it is expected that elevated temperatures will facilitate Ag diffusion to the surface, yielding a lubricious layer without wear of the hard matrix. The magnitude of this effect depends on the nature and quantity of the silver particles and their distribution within the coating matrix, which in turn are functions of the deposition parameters and the silver content of the film. Analogous to TiN/Ag coatings, incorporation of copper into the TiN film also results in a strong decrease of the friction coefficient, from high values of 0.6–0.7 corresponding to TiN films, to very low values of approximately 0.2. The hardness of the films ranges from 20 to 30 GPa (Musil and Vlček 2001). This means that the hardness of the TiN/Cu films is fully comparable with that of hard single-phase TiN films.

In another example, nanocrystalline yttrium-stabilized ZrO₂ (YSZ) grains encapsulated in a mixed YSZ-Au amorphous matrix showed interesting applications in aerospace (Voevodin et al. 2001). In this case, the large fraction of amorphous YSZ-Au grain boundary phase provided ductility by activating grain boundary slip and crack termination by nanocrack splitting. This provided a unique combination of high hardness and toughness in these coatings. The coating hardness was quite high, ranging from 18 to 30 GPa, while low friction coefficient of about 0.2 and high wear resistance are expected. Recently, this concept was expanded to high temperatures, where DLC and/or MoS₂ was combined with Au, providing high-temperature/low-temperature and dry/wet lubrication, and embedded in an yttria-stabilized zirconia (YSZ) matrix. Furthermore, advanced state-of-the-art

structures are designed to combine these composites with buried diffusion barrier layers and achieve surface self-adaptation during repeated temperature cycling. Recently, novel wear-resistant materials have been developed that combine nanocrystalline carbides (TiC, WC), oxide-based ceramics (YSZ and AlON), dichalcogenides (MoS₂, WS₂), and amorphous diamond-like carbon (DLC) into nanocomposite structures. Self-adapted coatings made of amorphous diamond-like carbon (DLC) matrix with incorporation of nanocrystalline TiC, WC, WS₂, and laser-processed MoS₂ reservoirs have demonstrated an order of magnitude improvement in toughness above that of single-phase carbides while maintaining the same level of hardness, a low friction coefficient in cycling from dry to humid environments, and an extremely long life in both ambient and space environments (Voevodin and Zabinski 2005). The surface chemistry, structure, and mechanical behavior of these nanocomposite materials are shown to reversibly change in the tribological contact, depending on applied loads and operational environment to maintain low friction and prevent wear.

Novel nanocomposite designs for self-lubricating hard coatings are very promising and provide a very attractive alternative to multilayer architectures. Nanocomposite coatings are more easily implemented, since they do not require precise control in the layer thickness and frequent cycling of the deposition parameters, as is required for fabrication of multilayer coatings. They are, however, relatively recent developments, and suitable scale-up of deposition techniques is currently under intense study.

Key Applications

Self-lubricating hard coatings with low friction coefficients and relatively high hardness have potential applications in industry. Their development will enable increased utilization of coatings in many types of applications including those in the automotive, tool, and aerospace industries where the operational environment is variable and severe. The typical use of advanced self-lubricating hard coatings has been proven to be very effective in improving the tribological performance of components such as cutting tools and ball bearings. On the one hand, requirements for personnel safety and low environmental impact in metal cutting/forming technologies today require using reduced amounts of or, in many cases, even no cooling/lubricating fluids. On the other hand, the requirements for low cost and high productivity set the necessity of using high-speed tools. Combining these two harsh demands has triggered a serious demands on the field of the self-lubricating hard wear-resistant coatings as a means to significantly reduce tool wear and friction in

dry cutting applications. There is also an increasing interest in the development of these advanced self-lubricating hard coatings that could provide low friction in vacuum environments, such as the applications in aerospace assemblies. Soft coatings commonly used in vacuum environment lubrication include dichalcogenides such as molybdenum disulfide and tungsten disulfide, and soft metals such as silver (Ag) or gold (Au). The incorporation of a soft metal within a hard transition metal carbide or nitride may therefore provide both high wear resistance and low friction in vacuum, even in a repeatedly switching atmosphere process.

Cross-References

- [Diamond-Like Carbon Coatings](#)
- [Doped MoS₂ Coatings and Their Tribology](#)
- [High-Temperature Solid Lubricating Materials](#)
- [Solid Lubricants, Ceramic-Based Self-lubricating Materials](#)

References

- T. Aizawa, T. Akhadejdamrong, A. Mitsuo, Self-lubrication of nitride ceramic coating by the chlorine ion implantation. *Surf. Coat. Technol.* **177–178**, 573–581 (2004)
- E. Badisch, G.A. Fontalvo, C. Mitterer, The response of PACVD TiN coatings to tribological tests with different counterparts. *Wear* **256**, 95–99 (2004)
- C. Donnet, A. Erdemir, *Tribology of Diamond-Like Carbon Films – Fundamentals and Applications* (Springer, New York, 2008)
- A. Erdemir, R.A. Erck, J. Robles, Relationship of Hertzian contact pressure to friction behavior of self-lubricating boric acid films. *Surf. Coat. Technol.* **49**, 435–438 (1991)
- S. Hogmark, S. Jacobson, M. Larsson, Design and evaluation of tribological coatings. *Wear* **246**, 20–33 (2000)
- P.H. Mayrhofer, P.Eh. Hovsepian, C. Mitterer, W.-D. Münz, Calorimetric evidence for frictional selfadaptation of TiAlN/VN superlattice coatings. *Surf. Coat. Technol.* **177–178**, 341–347 (2004)
- P.H. Mayrhofer, C. Mitterer, L. Hultman, H. Clemens, Microstructural design of hard coatings. *Prog. Mater. Sci.* **51**, 1032–1114 (2006)
- C.P. Mulligan, D. Gall, CrN–Ag self-lubricating hard coatings. *Surf. Coat. Technol.* **200**, 1495–1500 (2005)
- J. Musil, J. Vlček, Magnetron sputtering of hard nanocomposite coatings and their properties. *Surf. Coat. Technol.* **142–144**, 557–566 (2001)
- J. Robertson, Diamond-like amorphous carbon. *Mate. Sci. Eng. R* **37**, 129–281 (2002)
- D.G. Teer, New solid lubricant coatings. *Wear* **251**, 1068–1074 (2001)
- A.A. Voevodin, J.S. Zabinski, Nanocomposite and nanostructured tribological materials for space applications. *Compos. Sci. Technol.* **65**, 741–748 (2005)
- A.A. Voevodin, J.G. Jones, J.J. Hu, T.A. Fitz, J.S. Zabinski, Growth and structural characterization of yttria stabilized zirconia-gold nanocomposite films with improved toughness. *Thin Solid Films* **401**, 187–195 (2001)
- M. Woydt, A. Skopp, I. Dörfel, K. Witke, Wear engineering oxides/antiwear oxides. *Wear* **218**, 84–95 (1998)

Self-lubricating Metal Composite Coatings by Electrodeposition or Electroless Deposition

C. T. JOHN LOW, FRANK C. WALSH

Faculty of Engineering and the Environment, Engineering Sciences, University of Southampton, Southampton, UK

Synonyms

[Electrochemical deposition for self-lubricating metal composite coatings](#); [Electroplating for self-lubricating metal composite coatings](#)

Definition

Electrodeposition or electroless deposition is used to prepare a self-lubricating metal composite coating that offers an in situ lubricating film without the need for an external supply of lubricant. This composite coating consists of at least two constituent parts, namely a metal matrix and fine particles of lubricant dispersed throughout the metal matrix.

Scientific Fundamentals

Background

Lubrication is a process used to reduce friction and wear between two surfaces in close proximity. The two surfaces can either be stationary or in motion relative to each other. The most effective way to reduce friction and wear is to separate the two surfaces by means of a third body in the form of a lubricating film. The lubricant serves the purpose of limiting contact pressure, reducing friction and wear, and preventing galling and seizure between two surfaces. This lubricating film can be in the form of a solid lubricant, a mixture of solid/liquid dispersion, or a liquid phase lubricant.

The introduction of a third body lubricating film between two surfaces may not be adequate in a number of applications, e.g., liquid lubricant is not suitable for applications in a vacuum environment or aerospace applications, at extremely high contact pressure, or at elevated service temperature. Solid lubricants would be a more appropriate choice for these cases and in dry environment applications, but they have a finite service life depending on the type, size, and dimension of the lubricant. Replenishment of solid lubricants may be possible but this approach is not always practical during service.

One method for providing a continuous supply of lubricant is to incorporate the lubricant into a metallic coating matrix. This is known as a self-lubricating metal composite coating since the lubricant can be self-released to lubricate surfaces during service, which offers an in situ lubricating film without the need for an external supply of lubricant. Electrochemical wet processing techniques such as electrodeposition or electroless deposition can be used to prepare self-lubricating metal composite coatings. The metal composite coating contains at least two constituents parts – a metal matrix and fine particles of lubricant dispersed throughout the metal matrix.

Deposition of Metal Composite Coating

Electrodeposition or electroless deposition of metal composite coating is an electrochemical process carried out in a liquid solution under atmospheric pressure (Low et al. 2006). The liquid solution, also known as electrolyte or bath, is electrolytically conductive and contains:

1. A source of the metal to be deposited, i.e., dissolved metal salts in solution leading to metal ions,
2. Lubricant in the form of fine particles suspended in the solution, and
3. A variety of electrolyte additives, e.g., reducing agents, levelers, brighteners, wetters, and stress modifiers necessary to produce a good-quality coating.

Electrodeposition is an electrolytic process that uses the supply of an external electrical current to reduce metal salt in a solution to form metal. Electroless deposition is an auto-catalytic process that uses reducing agent in the solution to reduce metal salt to metal via catalytic chemical reactions. In both deposition processes, the self-lubricating metal composite coating is formed as the lubricating particles are co-deposited into the growing matrix of the deposited metal coating.

The embedded lubricant can be in the form of solid lubricant or liquid lubricant encapsulated in a hollow casing. When two surfaces come into contact, the embedded lubricant can be self-released to its surroundings, as a result of coating wear, abrasion, or friction forces, to lubricate surfaces. The coating can be deposited to several hundreds of microns thick, and the lubricant can be nanosized to several microns. To take advantage of the self-lubricating feature, the metal composite coating should meet certain requirements:

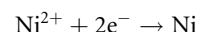
1. The metal coating should contain an even distribution of lubricating particles over the entire geometry of the coated object, to provide uniform coating properties.

2. The lubricating particles should disperse homogeneously throughout the coating layer thickness, or in a gradient composition for functional release purpose.
3. The metal composite coating should maintain a continuous surface lubricity during service and low shear strength, where the lubricating particles can be readily exposed to the service environment to combat tribological issues.
4. The choice of a metal matrix and lubricating particles should be compatible, to minimize issues, e.g., metal oxidation or corrosion, undesirable loss of lubricating particles or formation of unwanted products.
5. The metal composite coating should maintain a good adhesion to substrate and minimal surface cracks and coating stress, to provide a durable coating.

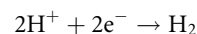
Principles of Electrodeposition

Figure 1 shows a pictorial representation of electrodeposition of a nickel composite coating. An external power source provides the supply of an electrical current. It connects the two electrodes, cathode and anode, which are immersed in a conductive solution containing nickel ions and lubricating particles. Nickel coating is electrodeposited on the surface of cathode, i.e., the work piece. The lubricating particles can be transported to the cathode via mechanical agitation or electrochemically via surface charged particles or by surfactants. Electrons follow from the anode to the cathode via an external electrical circuit.

At the cathode, electrodeposition of nickel is the main reaction:

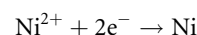


Hydrogen evolution may occur as a secondary reaction:

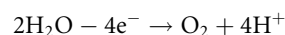


At the anode, the electrochemical reaction can be the dissolution of a nickel metal at a soluble anode, which has the purpose of maintaining a constant concentration of nickel ions in solution or at an insoluble anode, e.g., platinum, where oxygen evolution reaction occurs:

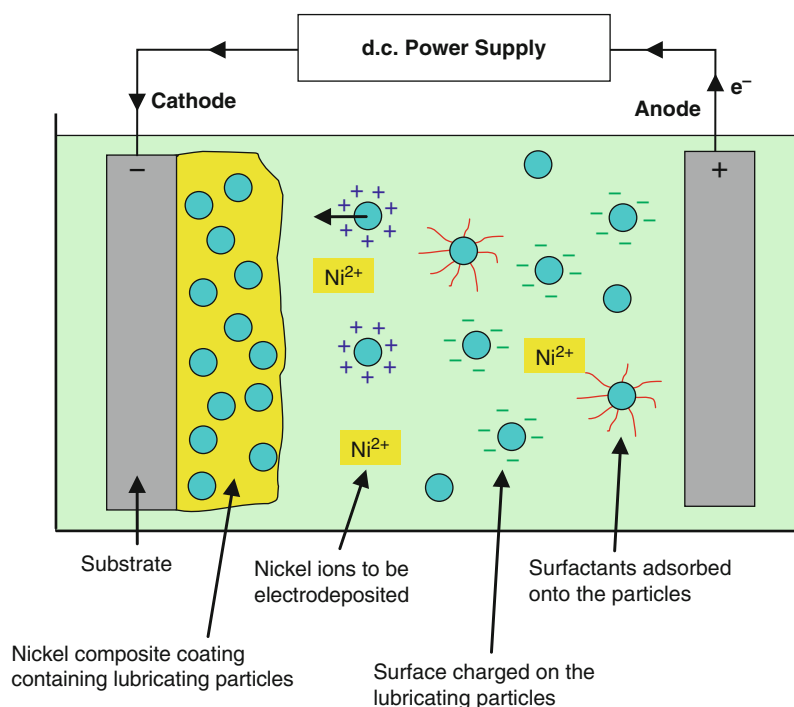
Soluble nickel anode



Insoluble platinum anode



In an acidic solution, the hydrogen evolution occurring at the cathode will result in a localized increase in pH



Self-lubricating Metal Composite Coatings by Electrodeposition or Electroless Deposition, Fig. 1 Pictorial presentation of the electrodeposition of nickel composite coating

near the surface of the electrode. It can lead to a reduction in the current efficiency of the system. Boric acid is usually added to the solution as a pH buffer to stop the cathode from becoming too alkaline. Nickel chloride is also added to assist the dissolution of nickel anode, without this, the nickel anode may be passivated.

The successful inclusion of lubricating particles into a metal coating is dependent on many process parameters, including the applied current waveform (e.g., DC constant current, pulsating current, duty cycle, and frequency), electrolyte composition (e.g., pH, concentration of metal ions, additives, temperature, conductivity, and viscosity), particle characteristics (e.g., surface charge, type, size, and shape), fluid flow in the solution (e.g., agitation of the bath or electrode movement), and geometry of the electrodeposition tank. The chemical and physical properties of deposited coating will depend on the solution composition, which in turn, depends on its formulation and operating conditions of the deposition process.

Many of these operating parameters are interrelated, and process optimization requires an understanding of the correlation of coating properties with their reaction environments. It is necessary to develop a stable colloidal solution, where the lubricating particles are

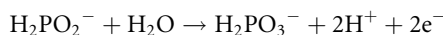
homogeneously mixed with metal salts and can be operated for many bath turnovers. The particles should be readily transported to the cathode surface and uniformly distributed, without agglomeration, into the growing metal matrix. Surface preparation of the particles is important to achieve this.

One approach includes the deployment of surfactant technology to modify the surface charge of particles, to favor its transportation to the cathode and maintain a stable dispersion in the solution. Types of surfactants may include non-ionic, anionic, cationic, and amphoteric. The solution typically contains a mixture of different types of surfactants. Zeta-potential can be used to provide information about the surface charge of particles in solution. Another method may include chemical preparation of the particles with reducing/oxidizing agents, to remove surface oxides and contaminants. It is also important to manipulate the hydrophobic/hydrophilic properties of the particles to reduce surface tension and allow a better wetting to the metal matrix. Inorganic additives are preferred to modify the surface state of particles since the use of organic additives can cause instability of the solution, i.e., decomposition of additives, leading to a high stress or brittle coating.

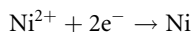
Principles of Electroless Deposition

Figure 2 provides a pictorial representation of an electroless deposition of nickel composite coating. This process uses only one electrode, i.e., the work piece. The solution needs to contain a reducing agent, e.g., hypophosphite ion, to react with nickel ions in solution to deposit nickel. This is an electroless deposition reaction since the reaction, i.e., chemical reduction of metal ions with reducing agent occurs in the absence of an external applied electrical current. The cathodic reduction of nickel ions occurs at a catalytically active surface. The electroless deposition is also commonly known as an auto-catalytic deposition reaction, which is driven by the electrons freed from the oxidation of a reducing agent.

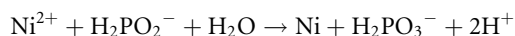
The anodic reaction is the oxidation of hypophosphite ion to orthophosphate ion:



The cathodic reaction is the deposition of nickel:



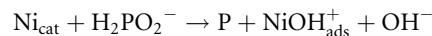
Overall reaction:



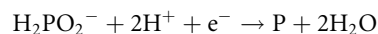
An alloy of Ni-P coating is usually deposited. The level of phosphorus in the coating affects the hardness and corrosion properties. Higher phosphorus coatings, e.g., >13 wt% are non-magnetic and have good corrosion resistance but are softer than the low phosphorus coatings. The lower phosphorus coatings, e.g., <9 wt% are harder

and have an improved wear resistance but not quite as resistant to corrosion. A number of co-deposition mechanisms of phosphorus have been postulated; some examples follow:

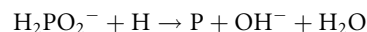
Direct interaction of a catalytic nickel surface, Ni_{cat} , with hypophosphite ion, where OH^- represents hydroxyl ions and $\text{NiOH}_{\text{ads}}^+$ represents a hydrolyzed Ni_{ads} species adsorbed at the catalytic surface:



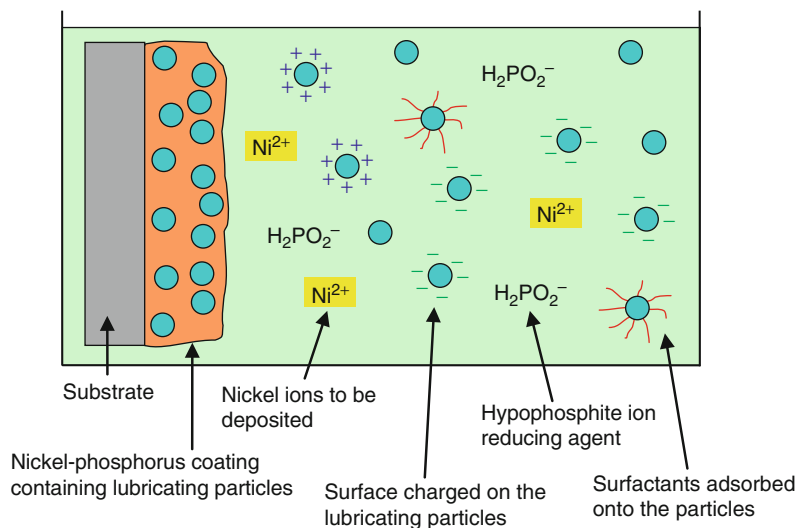
Cathodic electrochemical reduction of hypophosphite ion to produce elemental phosphorus and water:



Chemical reduction of hypophosphite ion by adsorbed atomic hydrogen on the catalytic surface to yield phosphorus, water, and hydroxyl ion:



In electrodeposition, the homogeneity of coating layer thickness will depend on the distribution of local current density over the entire surface of the work piece. Non-uniform current distribution often produces varying coating thickness in different regions. To achieve the required coating thickness in low current, density region frequently results in excessive thickness in the high current density region. Electroless deposited coatings are produced catalytically, meaning that the surfaces over the entire work piece can be deposited uniformly, i.e., non-line-of-sight deposition. Uniform coating thickness is produced



Self-lubricating Metal Composite Coatings by Electrodeposition or Electroless Deposition, Fig. 2 Pictorial presentation of the electroless deposition of nickel composite coating

wherever the work piece is in contact with the solution, regardless of the geometry of the coated object. Many post-plating finishing steps, such as grinding and mechanical polishing, to achieve the desired dimensional tolerance, can be eliminated in the electroless process.

Because electroless deposition is a chemical reduction process, both the metal ions and reducing agents will be consumed during operation, and thus a continuous replenishment of the solution with fresh stock is necessary to maintain a constant composition together with an optimum solution pH, conductivity, and viscosity. Various stabilizers are added to the solution to regulate the speed of metal deposition and minimize decomposition of the solution, which is inherently unstable. It is necessary to filter out the oxidized waste products generated to maintain a clean and long-lasting solution. Process control is therefore an important aspect of electroless deposition.

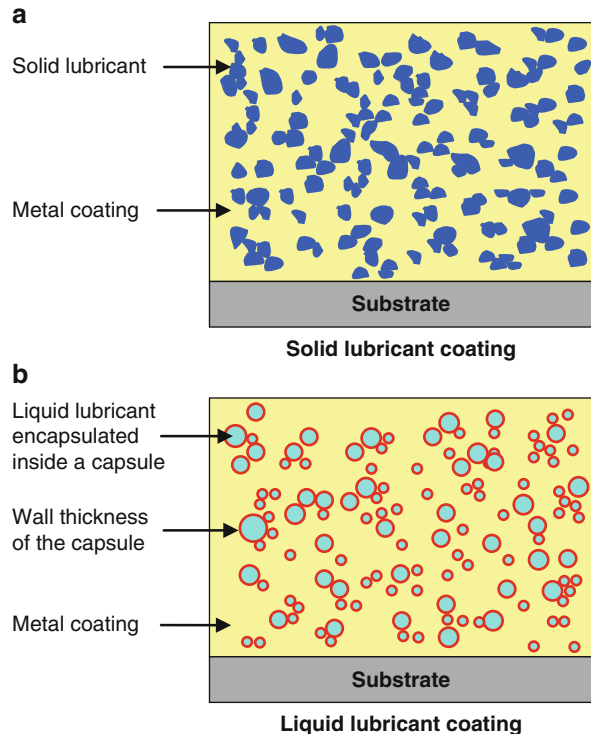
Transportation of particles to the work piece in the electroless deposition process can be similar to electrodeposition. Because this occurs in the absence of an electrical current in the electroless process, the surface properties of the particles play an even more crucial role in providing a successful incorporation into the metal coating. Factors such as surface charge, hydrophobic/hydrophilic nature, and size/type/dimensions of the particles will influence the process. Other approaches may include mechanically transporting the particles to the work piece via rotating barrel, up-down paddle, or reciprocating jig holder connected to an external motor, air-pump installed with the tank to generate air bubbles to keep the particles in suspension, or ultrasonic wave power in solution to provide less agglomerated particles.

Types of Self-lubricating Metal Composite Coatings

Figure 3 provides a pictorial representation of the two different types of self-lubricating metal composite coatings, categorized according to the nature of lubricant. In both cases, the metal coating provides a matrix to bind the lubricant in place. The coatings can be:

- (a) Dry-based lubrication via solid lubricant
- (b) Wet-based lubrication via liquid lubricant encapsulated in a capsule

In the solid lubricant coating, the self-lubricating property comes into service when the two opposing surfaces (or coatings) are in contact. As the coating wears, the solid lubricant near the top surface provides a dry lubricity environment for the other opposing surface. This provides a low shear surface, minimizes the contact pressure spots, and protects the surfaces from excessive wear at high load



Self-lubricating Metal Composite Coatings by Electrodeposition or Electroless Deposition, Fig. 3 Pictorial representation of the two different types of self-lubricating metal composite coatings. (a) Dry-based lubrication via solid lubricant and (b) wet-based lubrication via capsules filled with liquid lubricant

applications. As the coating continues to wear under friction, the solid lubricant can be gradually exfoliated to expose the service environment. A continuous supply of solid lubricant to lubricate surfaces in situ is possible in a coating containing solid lubricant embedded throughout the layer thickness.

In the liquid lubricant coating, the lubricant is in the form of liquid, which is encapsulated inside a hollow capsule. The self-lubricating property occurs when the capsule is ruptured mechanically under friction or shear force, which releases the liquid lubricant. This type of coating provides a wet-phase lubricity surroundings, which finds applications in low contact pressure, low load, or low temperature environments. Most capsules are spherical, with diameters ranging from several tens of nanometers to a few micrometers. The capsule wall can be made from a synthetic polymer, e.g., polyvinyl alcohol, or other types such as cellulose or gelatine. The capsule wall does not have lubricity properties. Once the liquid

lubricant is released, the capsule wall either remains embedded with the coating or wears off during service.

In addition to liquid lubricant-filled capsules, other functional liquid compounds can be encapsulated to provide a multi-functional coating. For example, liquid-form corrosion inhibitors or liquid fillers can be embedded into the coating, which has the purpose of protecting against localized corrosion and self-repairing damage sites. Many methods can be used to prepare the liquid-filled capsules, including interfacial polymerization, in-situ polymerization, solvent evaporation, and phase separation in an aqueous/organic solvent. Typically, the liquid lubricant, e.g., oil, is emulsified in a solution and then stabilized in another immiscible solution that will eventually form the capsule wall. It is desirable to form capsules with a narrow size distribution, less agglomeration, and a controlled wall thickness. Selection of wall materials, stabilizers, and the emulsification process is dependent on the type of liquid to be encapsulated. Care must be taken to limit the liquid reacts adversely with the metal.

The density of lubricating particles, both solid and liquid type, embedded in the metal coating varies and is highly dependent on the particles sizes and dimensions together with the deposition conditions. Particle size can be tens of nanometer to several microns, and in the form of irregularly shaped particles or controlled dimension nanotubes, nanowires, or nanospheroids (Low et al. 2010). The co-deposition of nanosized or submicron particles with metal matrices demonstrates superior coating properties compared with micron sized particles. The incorporation of nanosized particles is challenging due to their high surface energy and small size. An understanding of the particles' interaction with the reaction environment, e.g., electrolyte composition and process conditions, is mandatory to improving coating properties.

The choice of particle density is important. Typically, a particle density of 20–30 vol% is suitable for most wear applications. The concentration of particles in the coating should be adjusted to provide optimum performance. Too low particle density may not provide sufficient properties derived from the particles; and too high particle density may lead to premature detachment of the particles as the coating lacks wear resistance and mechanical integrity, i.e., the metal matrix is not able to tightly bind all the particles in place. In many cases, the coefficient of friction decreases when the volume content of the lubricant is increased. The service lifetime of a coating will also depend on the coating thickness. Thick coatings can provide a longer service lifetime but should not be too thick, which could alter the shape and dimension of a coated object. A thin coating that is more durable and has a better performance than

a thick coating is favorable but should not be too thin, which could result in a shorter wear life.

Key Applications

Self-lubricating metal composite coatings provide a continuous in situ lubricating film to combat surface tribological issues. The coating may perform functions:

- (a) Reduce surface friction to minimize wear
- (b) Keep moving parts separated
- (c) Prevent galling or seizure between two surfaces
- (d) Prevent corrosion or surface oxidation
- (e) Transfer heat or carry away debris, and
- (f) Prevent surface fouling.

Key applications of the coatings can be seen in load-carrying components or for film transfer lubrication. The most common examples are in the operation of mechanical systems such as gears and traction devices, turbines, ball bearings, roller bearing retainers, piston rings, and cylinder liners in engines. Without lubrication, the pressure between two surfaces would generate enough heat to rapidly damage the surfaces, which could lead to seizure of the components. Applications are also seen in sliding electrical contacts and brushes, mold release components, and anti-fouling heat exchange, household heat elements, and flow pipes to prevent buildup of lime scale or unwanted accumulation of materials.

Self-lubricating coatings are also used in machining applications and metal cutting tools such as milling, drilling, tapping, and stamping. They can be used together with an externally supplied lubricant or liquid system to cool the contact areas and remove wear products. Another application is in fail-safe operations, where the components, e.g., engine bearings, are able to continue to operate for an extended limited time following failure of the normal lubrication system. Self-lubricating coatings are also used extensively to coat the surfaces of grooves, slots, threads, nuts, bolts, and fasteners that can be easily tightened and unscrewed after a period of service.

The most commonly used metal matrix is nickel and its alloys, e.g., nickel-phosphorus or nickel-cobalt, due to its inherently good corrosion and wear resistance. Tin, copper, silver, precious metals such as platinum, palladium, or gold, and cobalt alloys such as cobalt-phosphorus or cobalt-tungsten are also used. The choice of a coating matrix is important. For example, soft noble metals, e.g., Ag, Au, and Pt, possess good thermal stability at elevated temperatures, i.e., due to their high melting points, but may suffer from a high coating wear as a result of the soft nature of the metal. These coating matrices can easily undergo plastic deformation and plowing during

sliding, which can lead to increased friction and irregular surface topographies. Hard coatings such as Ni–Co and Co–W have good coating toughness and resist wear better but may suffer from high coating stress, which can lead to coating delamination and poor adhesion to the work piece (Wang et al. 2005a, b, c, 2006).

Nanocrystalline or amorphous coatings are often deployed to offer enhanced coating properties compared to polycrystalline coatings. Nanocrystalline coatings typically show improved coating hardness and a large volume of grain boundaries, which helps to restrict the growth of crack size and to deflect or terminate growing cracks. Amorphous coatings can facilitate an improved grain boundary sliding leading to enhanced ductility and can prevent fracture under high load. The metal composite coating can also be heat treated in air or in a gaseous environment under vacuum to form new phases or intermetallics and to improve the coating properties, e.g., enhance hardness, reduce coating stress, and improve coating adhesion.

Electroless Ni-PTFE coating garners the most commercial interest. The coating can typically contain 10–25 vol% PTFE. It is widely deposited and fairly inexpensive but has a limited high temperature application since PTFE decomposes at temperature above 300°C. It is a soft coating due to the soft nature of PTFE particles and is commonly used for lower temperature and light loading applications. Inorganic lubricants are harder and can withstand higher service temperatures and are becoming increasingly popular. Examples include hexagonal boron nitride (h-BN), sulfide lubricants such as molybdenum disulfide (MoS₂), and tungsten disulfide (WS₂). The lubricating properties of these solid lubricants are attributed to a lamellar layered structure with weak bonding between layers. Such layers are able to slide relatively parallel to each other with minimal force, thus giving them low friction properties and a low shear stress region suitable for high load application. Carbons and graphite are commonly used, and carbon nanotubes, both single walled and multilayered, also have lubricating properties.

Each solid lubricant has a specific operation temperature at which its lubrication mechanism can work. For example, graphite can provide lubrication up to 400°C but needs a humidified environment to facilitate lubrication. h-BN is a ceramic lubricant with a high thermal conductivity and can withstand temperatures up to 1,000°C. MoS₂ has a hexagonal crystal structure with the intrinsic property of easy shear, but it has poor resistance to oxidation, e.g., it can be used up to 350°C in air, but 1,100°C in reducing environments.

Although these solid lubricants can provide lubricity at high temperature, many metal composite coatings

usually degrade in coating performance due to thermal instability of the metal. Nanostructured metals, alloys, and multilayered or functionally graded coatings that can withstand high temperature environments and, for cyclic temperature application, can be used to extend the service lifetime of a coating. The selection of a suitable coating for tribological applications will depend on the types of lubricant and the properties of metal composite coating, which in turn depends on the application environment and service conditions.

Cross-References

- [Electro- and Electroless Composite Coatings](#)
- [Electrochemical Deposition](#)
- [Electroplating](#)

References

- C.T.J. Low, R.G.A. Wills, F.C. Walsh, Electrodeposition of composite coatings containing nanoparticles in a metal deposit. *Surf. Coat. Technol.* **201**(1–2), 371–383 (2006)
- C.T.J. Low, J.O. Bello, J.A. Wharton, R.J.K. Wood, K.R. Stokes, F.C. Walsh, Electrodeposition and tribological characterisation of nickel nanocomposite coatings reinforced with nanotubular titanates. *Surf. Coat. Technol.* **205**(7), 1856–1863 (2010)
- L.P. Wang, Y. Gao, Q. Xue, H. Liu, T. Xu, Microstructure and tribological properties of electrodeposited Ni–Co alloy deposits. *Appl. Surf. Sci.* **242**, 326–332 (2005a)
- L.P. Wang, Y. Gao, H. Liu, Q. Xue, T. Xu, Effects of bivalent Co ion on the co-deposition of nickel and nano-diamond particles. *Surf. Coat. Technol.* **191**, 1–6 (2005b)
- L.P. Wang, Y. Gao, Q. Xue, H. Liu, T. Xu, Effects of nano-diamond particles on the structure and tribological property of Ni-matrix nanocomposite coatings. *Mater. Sci. Eng. A* **390**, 313–318 (2005c)
- L.P. Wang, J. Zhang, Z. Zeng, Y. Lin, L. Hu, Q. Xue, Fabrication of a nanocrystalline Ni–Co/CoO functionally graded layer with excellent electrochemical corrosion and tribological performance. *Nanotechnology* **17**, 4614–4623 (2006)

Self-Lubricating Treatment of Light Alloys

RAINER GADOW, DIETMAR SCHERER

Institute for Manufacturing Technologies of Ceramic Components and Composites, University of Stuttgart, Stuttgart, Germany

Synonyms

[Metall-, cermet- or ceramic-polymer-composite coatings for tribological applications](#); [Self-lubricating coatings and composite coatings](#)

Definition

Self-lubricating treatment of light metal alloys contains different techniques to improve the wear and friction behavior of machine and vehicle components made of light alloys.

Introduction

In mechanical engineering there is an increasing demand for lightweight design and materials engineering. Light alloys including magnesium (Mg), aluminum (Al), and titanium (Ti) are widely used in for aerospace components, machine elements, and instruments as well as in automobile parts; they offer weight reduction, reduction of consumed energy, and upgrading of performance. One major drawback of light metal alloys is their poor surface properties when it comes to friction and wear. Thus, for successful implementation of these materials under high surface loadings, a suitable and well-designed surface treatment or protective and functional surface coating is necessary.

In applications with sliding contact surfaces, wear degradation occurs as a result of friction between the contacted components. The most common way to reduce friction and wear is the use of grease and lubricating oils. However, due to more stringent environmental requirements, the use of these classic lubricants will be limited in future because most lubricants contain a certain amount of ecologically harmful chemical additives. Therefore, the tribologically stressed machine elements will have to operate with a smaller amount of lubricants, life-time lubrication, or at least extended drain intervals. When liquid lubrication is either not permitted or fails to function, e.g., due to slow relative motion, advanced coating materials with solid lubricant ability will be used to provide low friction and wear coefficients.

A wide variety of functional metallurgical, ceramic, and cermet coating systems can be deposited on light metal substrates in an effective way by different thermal spray processes. These coatings provide hardness and compressive strength ability as well as creep and wear resistance. In order to improve the tribological performance of thermally sprayed coatings, either solid lubricants are incorporated in the coating during thermal spraying (Brune et al. 1996) or the coating material itself acts as a self-lubricating surface. For the latter approach, metal oxides from titanium, vanadium, molybdenum, and tungsten are suitable coating materials (Woydt et al. 1998). These so-called lubricious oxides indicate solid lubricant abilities due to special crystallographic shear plane structures in the lattice of the solid layer, forming a polycrystalline structure resulting in a low-friction coefficient.

Here the focus is on the tribological evaluation of thermally sprayed TiO_2 coatings under dry sliding or single oil shot conditions. In the first part of the experiments, TiO_2 coatings were applied by atmospheric plasma spraying (APS) with variation of the process parameters using different plasma torches. Successively, the crystal structure and phase composition of the titania coatings were X-ray analyzed and the tribological performance was determined depending on the selected process parameters. In the second part of the experiments, TiO_2 coatings were deposited on light metal substrates as primary layers and subsequently polymers containing microscale solid lubricant particles, so-called lubricant lacquers, as well as MoS_2 thin films with dry lubricating properties were applied and the tribological performance of these combined coatings was again evaluated under dry friction conditions.

Scientific Fundamentals

It is well-known that light alloys as monolithic materials show poor tribological performance due to their intrinsic properties. The main failure mode of light alloys in tribological contacts is adhesive wear, which leads to material transfer to the counterpart and so to a rising friction coefficient and finally to system failure.

There are different approaches to enhancing the tribological abilities of light alloys. Basically, there are two different methods to achieve a better frictional behavior; first, tribologically functional fillers in micro- and meso-scale can be deposited in light alloy matrices and second tribologically functional coatings as thin solid films can be applied on top of the light alloys. The first method is mainly focused on the improvement of the wear behavior and the structural properties, e.g., the stiffness.

To realize self-lubricating abilities, the only promising approach is covering the surface of the light alloy components with functional coatings. Depending on the application requirements, different coating techniques and coating materials can be used to modify the surface properties of the composite. Due to the required properties, especially under dry-sliding or deficient lubrication, only a few materials are able to meet the requirements.

Solid Lubrication

The most common way to improve the frictional behavior under dry conditions is to use solid lubricants. There are three different types of solid lubricants (Plagge 1991; Schneider 1987):

1. Structural lubricants (BN, $\text{C}_{\text{graphite}}$)
2. Mechanical lubricants (PTFE)
3. Chemical lubricants (organic or polymer compounds)

In addition to the so-called solid lubricants, there are several materials that show very interesting tribological properties. Most of these materials are based on metal oxides that show, due to their lattice structure, self-lubricating abilities comparable to solid lubricants.

Structural Lubricants

Structural lubricants are distinguished by their lattice structure. Their properties are mainly influenced by the anisotropic chemical bonding within the planes and between two planes. In the case of graphite and hexagonal boron nitride, the bonding strength within the planes is much higher than between the planes, and therefore a defined sliding of the planes is possible. The so-called shear plane structure gives these materials their typical friction behavior. The shear planes can easily slide parallel to the moving direction, and at the same time they are able to bear perpendicularly high compressive loads. The most common solid lubricant of this type is molybdenum disulfide (MoS_2). The lattice structure of MoS_2 consists of layers of S-Mo-S-planes (Fig. 1a). The bonding energy between the sulfide and molybdenum atoms is, due to the covalent bonds, much higher compared with the bonding energy between the sulfide atoms of two layers and therefore the shearing strength of the planes is quite low.

In comparison to MoS_2 , graphite and boron nitride show a quite different lattice structure (Fig. 1b). The atoms are ordered in hexagonal structures within the shear planes. The covalent bonds between the atoms within the hexagonal lattice planes lead to the high

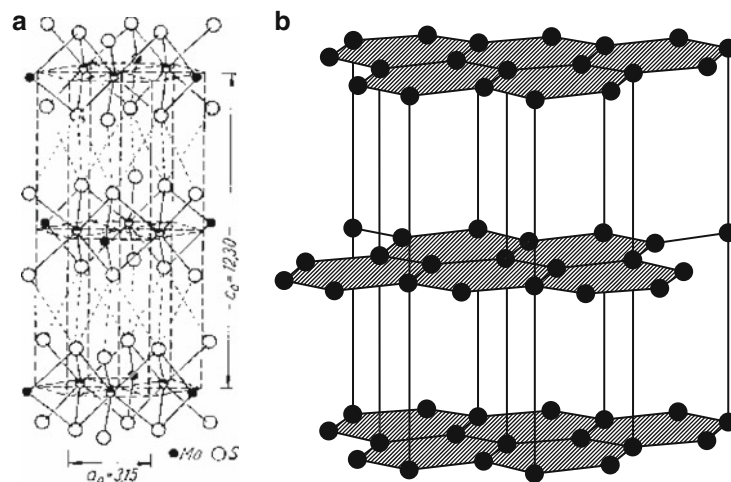
compressive strength perpendicular to the shear planes. The bond between two planes is based on van der Waals forces and, therefore, the bonding is much weaker and so the planes are able to slide parallel to the sliding direction.

Mechanical Lubricants

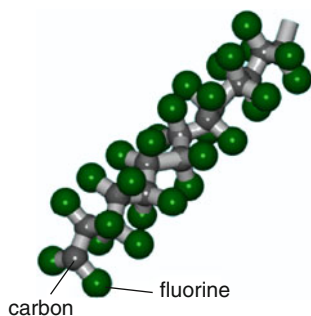
The active principle of mechanical lubricants isn't based on special lattice structures. The group of so-called white solid lubricants that have fluoropolymer origins are the most common mechanical lubricants. The self-lubricating abilities of these materials are based on the chemical bonding between the molecular chains. The most common agent of this lubricant type is polytetrafluoroethylene (PTFE). It is well known that PTFE is distinguished by its excellent non-stick effect and its low friction coefficient. The low wetting of PTFE is associated with its mechanical and tribological properties. The outstanding properties of PTFE are mainly caused by the bond character between the carbon-backbone and the fluorine (Fig. 2). Due to the strong bonding and the fact that fluorine isn't able to create more than one covalent bond, the intermolecular bonding is based on van der Waals forces. Thus, the self-lubricating ability of PTFE is based on weak intermolecular bonds, which offer the possibility of an easy slipping off between the molecular chains.

Chemical Lubricants

Chemical lubricants are usually used to coat iron-based components. These lubricants usually consist of chlorine-, fluorine-, sulfur-, or phosphor-based organic compounds. Due to chemical interactions between the surface and the



Self-Lubricating Treatment of Light Alloys, Fig. 1 Lattice structure of (a) Molybdenum disulfide (MoS_2) and (b) Graphite (Kleber 1967)



Self-Lubricating Treatment of Light Alloys, Fig. 2 Lattice structure of polytetrafluoroethylene (PTFE)

lubricant, a thin self-lubricating film is built up on top of the metal surface. These thin films can be worn out, but if liquid lubricant remains in contact the defects in the lubricant layer can be healed.

Self-Lubricating Metal Oxides: The System Ti-O – Magnéli Phases

The high potential of lightweight engineering can be effectively used by the application of special protective coatings on various light metal machine components. To enhance the tribological properties of light metal substrates, titania coatings or titania-based multilayer coatings can be applied using several coating techniques.

The physical and chemical properties of the titanium (Ti) and oxygen (O_x) compounds TiO_x are strongly influenced by the ratio of the single elements. The stoichiometrically correct ratio is TiO_2 , with a melting point of 1,855°C. In addition to the stoichiometric ratio, the system Ti-O contains a lot of phases with an oxygen deficiency. Depending on the degree of oxygen deficiency, the non-stoichiometry TiO_x phases show significant differences in hardness, color, electrical resistivity, tribological, and corrosive behavior. Phase compositions with an atomic ratio O/Ti between 1.6 and 2 are of technological significance and importance for tribological applications. This phase composition can be described by the chemical formula of TiO_{2n-1} ($4 < n < 10$), the so-called Magnéli phases. Magnéli first recognized that oxides of titanium and some other metal oxides form homologous series with planar stacking faults. Due to the oxygen defects in the lattice, a crystallographic shear plane structure occurs that enables a sliding of single-crystal layers (Anderson et al. 1967). As a result, these Magnéli phases show a low friction coefficient under tribological load. It is assumed that the slight non-stoichiometric phases possess the best solid

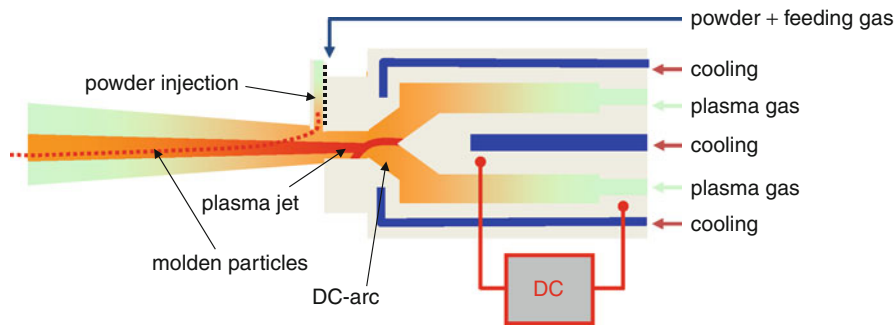
lubricant properties (Gardos and Gardos 1988). The tribological properties of these Magnéli phases have been investigated by several research teams.

Application of Tribologically Functional Coatings by Means of Thermal Spraying

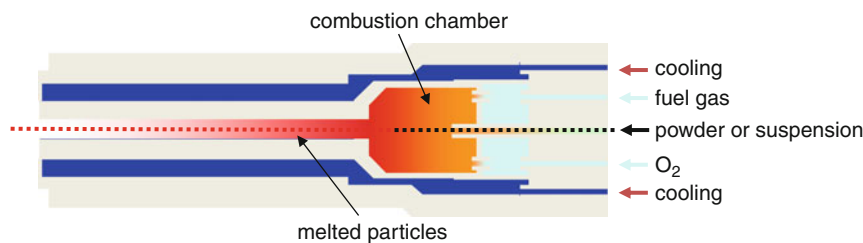
The thermal spray process offers the possibility to apply a broad variety of metallurgical, cermet, and ceramic coatings on the surface of machine components, even on systems with complex size and geometry. The coatings are commonly used to improve wear and corrosion resistance, operation temperature, and thermal shock behavior, or to influence the electrical, magnetic, and biological behavior of the layer composite surface.

Since about 1930, various thermal spray techniques have been developed. They are characterized by different temperature and acceleration rates as well as by the capability to handle different raw materials. Atmospheric plasma spraying (APS) is a cost-effective thermal spraying technology able to process all materials with a congruent melting behavior in an open system under atmospheric pressure conditions. The principle of an APS torch is shown in Fig. 3.

The energy is induced by a hot expanding plasma. An electric arc discharge between a water-cooled circular copper anode and a tungsten cathode dissociates and ionizes the working gas and builds up the plasma, which expands to the atmosphere, forming a jet. The plasma temperature can be as high as 20,000 K and the loaded powder particles can reach velocities up to 450 m/s. The powder, suspended in a carrier gas, is injected into the plasma. The particles are fully or partially melted and are accelerated versus the cold substrate. Upon impacting onto the surface, the fluid particles are rapidly solidified (quenched) and the coating is built up. Depending on the quenching rate, high temperature phases, and even partially amorphous coatings, can be realized. During the APS process temperatures up to 20,000°C are obtained, therefore this system is mainly used for refractory materials. In contrast, the high-velocity oxy fuel (HVOF) and high-velocity suspension flame spraying (HVSFS) processes represent an alternative coating technique. The HVOF process uses liquid fuels or fuel gases for high energetic combustion with oxygen ($v_{max} \sim 300\text{--}600$ m/s; $T_{max} \sim 2,500\text{--}3,200^\circ\text{C}$), Fig. 4. It leads to extremely dense coatings because of the high kinetic energy of the hot powder loaded gas jet. An additional benefit of the novel technique using HVSFS is the option of processing nanoscale powders and powder blends in suspensions broadening the possible material range and yielding in outstanding coatings.



Self-Lubricating Treatment of Light Alloys, Fig. 3 Principle of atmospheric plasma spraying



Self-Lubricating Treatment of Light Alloys, Fig. 4 Scheme of a HVOF/HVSFS – Torch

Before the coating process, grit blasting and degreasing of the substrate surface is performed. The roughening of the surface with corundum of defined size improves the mechanical adhesion of the coating and induces compressive stresses into the substrate material. Following the coating deposition, a mechanical (grinding, polishing, honing) or thermal post-treatment of the coating surface takes place.

Manufacturing of TiO_x Coatings by APS with Variation of the Process Parameters on Light Metal Substrates

The following experiments were performed using AlSi12 as substrate material. The goal of these experiments was to evaluate the influence of the selected plasma torch and hydrogen flow rate on the resulting mechanical and tribological properties of the thermally sprayed TiO₂ coating. Two different plasma torches were evaluated, a F4 torch for the coating of planar substrates and a F1 torch for the application of internal coatings. For both torches an Ar/H₂ plasma was used. The argon mass flow rate was kept constant and the hydrogen flow rate was varied between 3 and 15 l/min. For all experiments, a commercially available TiO₂ spray powder (Amperit 782.054) with a particle size distribution of $-45 + 10 \mu\text{m}$ was used. The deposition

of the TiO₂ coating on the planar AlSi12 substrates with the F4 torch was carried out using a meandering torch feed and a spraying distance of 100 mm. For the application of the internal TiO₂ coating on a rotating AlSi12 tube with the F1 plasma torch, a spraying distance of 30 mm was selected. During thermal spraying all substrates were simultaneously cooled by CO₂ and air. The resulting coating thickness for both processes was 200 μm .

The mechanical properties of the coatings deposited by the F4 and the F1 torch are summarized in Tables 2 and 3, respectively. The surface roughness of the TiO₂ coatings decreases with increasing hydrogen mass flow rate, because of the increasing total process mass flow rates as well as the higher process temperatures due to the increasing hydrogen content in the plasma flame. No significant correlation can be detected between hydrogen mass flow rate and bonding strength σ_B and the bonding strength is only slightly higher for the F1-coated samples. For the F4-coated samples, no strong dependence of the microhardness measured in the cross section HV0.05 C and the surface HV0.05 S on the hydrogen mass flow is visible, and the measured values on the surface are about 20% lower than the ones measured in the cross section (Table 1). For the F1-coated samples, the microhardness values measured on the surface and cross

Self-Lubricating Treatment of Light Alloys, Table 1

Selected mechanical properties of APS-F4 sprayed TiO_2 coatings on AlSi12 planar substrates

H_2 flow (l/min)	R_a	R_z	σ_B	HV0.05 _C	HV0.05 _S
5	4.02	23.36	17	922	783.5
10	3.67	20.35	14	935	785.5
15	3.05	19.86	15	858	785.9

Self-Lubricating Treatment of Light Alloys, Table 2

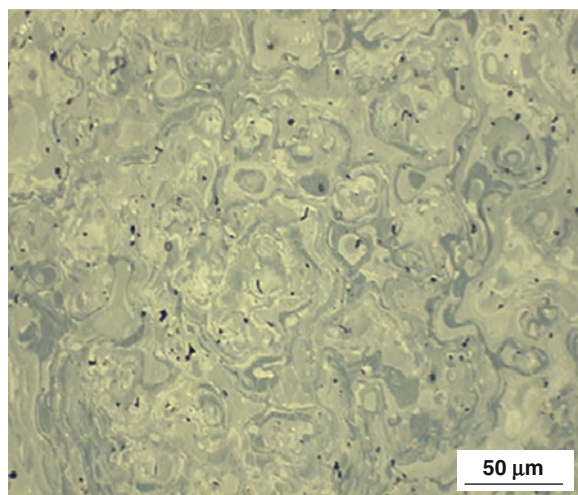
Selected mechanical properties of APS-F1 sprayed TiO_2 coatings on AlSi12 tubes

H_2 flow (l/min)	R_a	R_z	σ_B	HV0.05 _C	HV0.05 _S
3	12.61	73.62	21	1,239	867
6	11.62	67.04	20	1,177	824
9	5.41	36.60	25	949.5	650

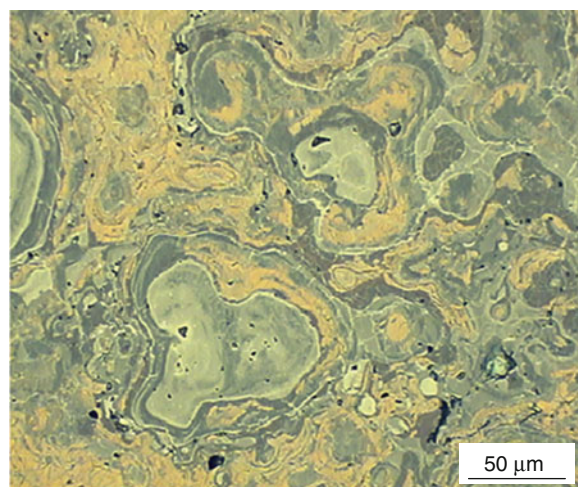
section decrease strongly with increasing hydrogen content (Table 2). In addition, deviations of about 30% between the measured hardness in the cross section and the coating surface can be detected.

For morphological investigations, microscopic shots of polished TiO_2 surfaces were recorded (compare Figs. 5 and 6.) Different colors of gray are visible on the polished coating surface. For the F1-coated samples, an additional brown-colored phase can be detected by polarized light microscopy. The amount of this brown-colored phase increases with increasing hydrogen mass flow rate. Because of the noticeable color inhomogeneities in the APS F1-sprayed TiO_2 coatings, hardness measurements were performed in the single phases, which reveal large differences in the microhardness (see Fig. 7). Using wavelength dispersive spectrometry (WDS) the Ti and O proportion in the phases was determined. For the dark (A) and bright (B) grayscale phases, a stoichiometry of pure TiO_2 and slight oxygen deficits can be measured. For the brown-colored phase (C), non-stoichiometries between Ti_4O_7 and Ti_7O_{13} were determined.

For the interpretation of the different gray and brownish colors, XRD measurements using a Siemens diffractometer D-500 were performed on the coating surfaces and compared to the TiO_2 spray powder (Figs. 8 and 9). The thermally sprayed TiO_2 coatings were analyzed in the

**Self-Lubricating Treatment of Light Alloys, Fig. 5**

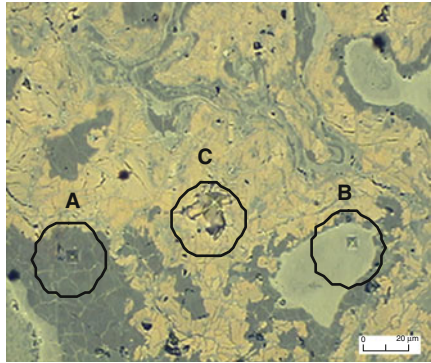
Microscope shots of the coating morphology – F4 torch, APS sprayed

**Self-Lubricating Treatment of Light Alloys, Fig. 6**

Microscope shots of the coating morphology – F1 torch, APS sprayed

state “as sprayed” using $\text{Cu K}\alpha$ radiation. The spray powder indicates ambiguous oxygen sub-stoichiometries in the measured XRD patterns (28° and 55°). A detailed sequence definition cannot be made because of the resolution limits of the diffractometry and evaluation software. The APS-F4 sprayed coating basically shows a pure TiO_2 stoichiometry with a predominant rutil crystal structure. The deviations in the gray color can be interpreted as slight deviations in the TiO_2 phase stoichiometry.

No deviations in the phase structure can be found dependent on the hydrogen content. The XRD results of the APS-F1 deposited coatings shown in Fig. 9 instead show in the range of 35–45° as well as around 55° large deviations



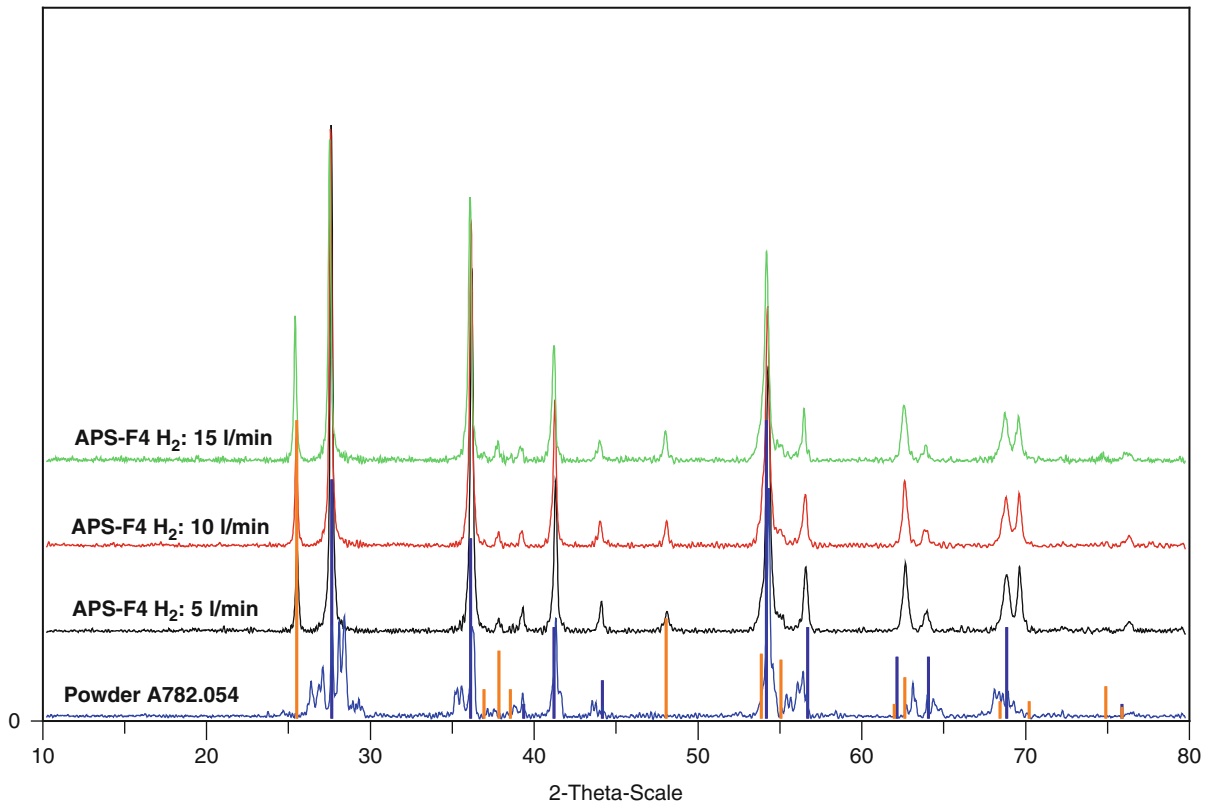
Microhardness
HV0.05
A: 1422,5
B: 1213
C: 622

Self-Lubricating Treatment of Light Alloys, Fig. 7
Microhardness HV0.05 in the single TiO_x phases – F1 torch, APS sprayed (H_2 : 9 l/min)

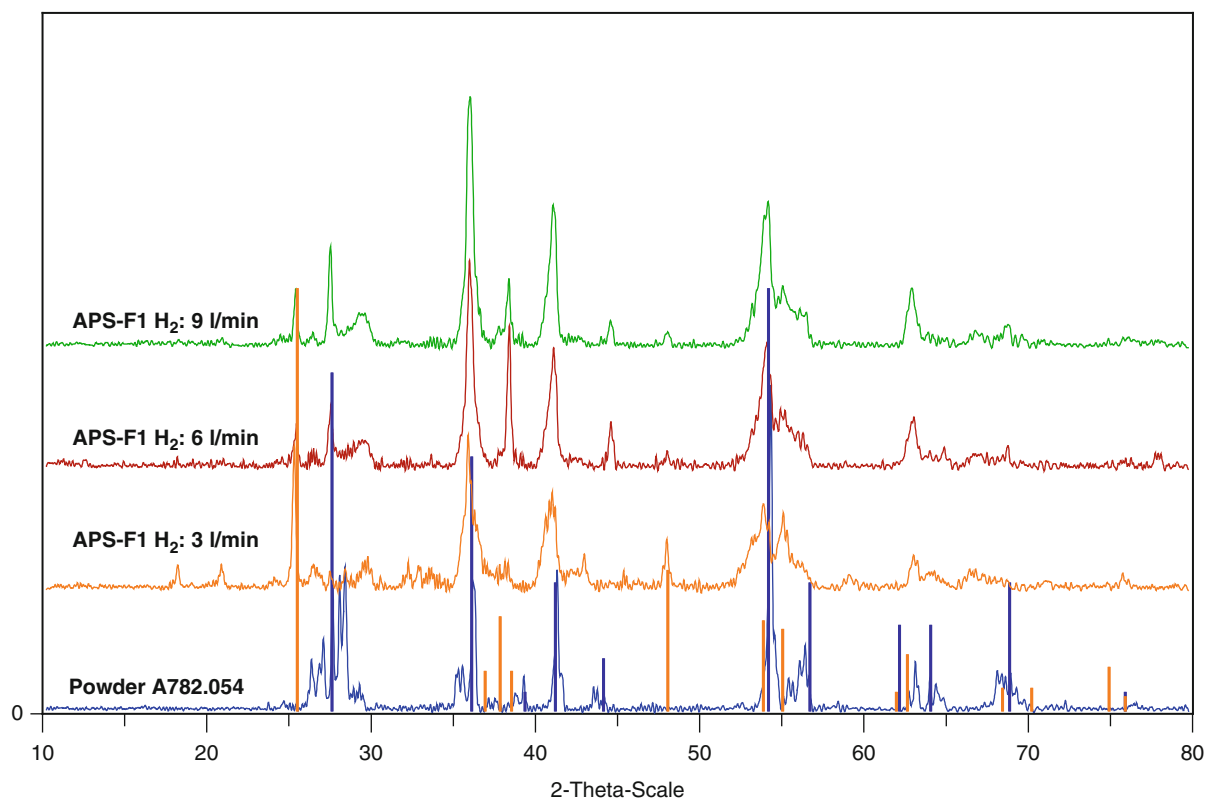
to the powder raw material as well as the APS-F4 deposited coating. These results indicate different oxygen sub-stoichiometries in the inhomogeneous APS F1 sprayed TiO_x coating structure.

After the mechanical and XRD characterization of the TiO_2 coatings, tribological investigations were performed under dry friction conditions using the same parameters as in the first test series. All coating surfaces were finished to an average surface roughness $R_a = 0.03 \mu\text{m}$. Figure 10 shows the measured wear and friction coefficients of the different TiO_2 coatings. For all APS-F4 sprayed coatings, high wear rates and friction coefficients were measured. For the APS-F1 sprayed coatings, decreasing wear and friction coefficients can be measured with increasing hydrogen mass flow rate.

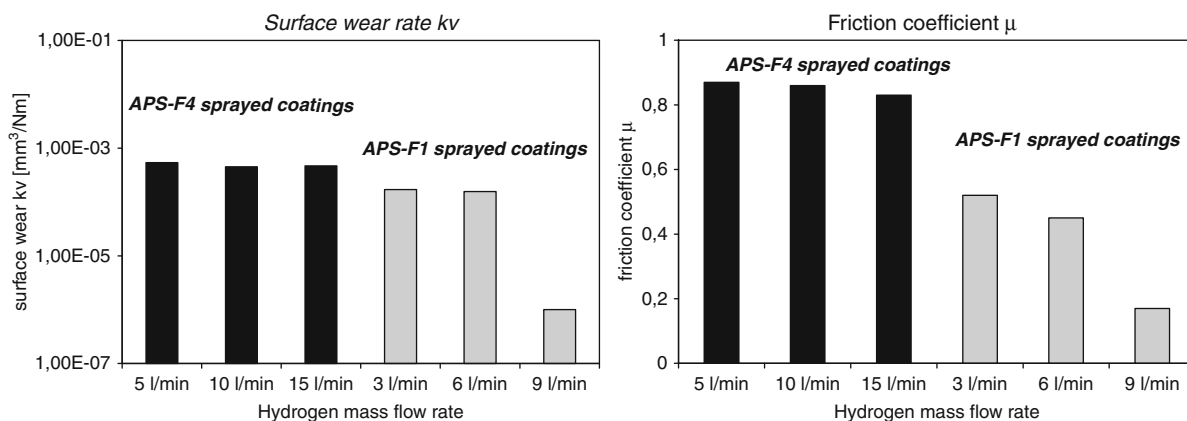
The high potential of lightweight engineering can be effectively used by the application of special protective coatings on various light metal machine components. Light metal substrates were coated by different APS plasma spray processes with TiO_2 coatings and TiO_2 -based multilayer coatings. The occurrence of tensile



Self-Lubricating Treatment of Light Alloys, Fig. 8 XRD measurement results of the APS- F4 torch sprayed coatings compared with the powder raw material



Self-Lubricating Treatment of Light Alloys, Fig. 9 XRD measurement results of the APS- F1 torch sprayed coatings compared with the powder raw material



Self-Lubricating Treatment of Light Alloys, Fig. 10 Surface wear rate and friction coefficient of the investigated TiO₂ coating systems under dry friction conditions for different plasma torches and hydrogen flow rates

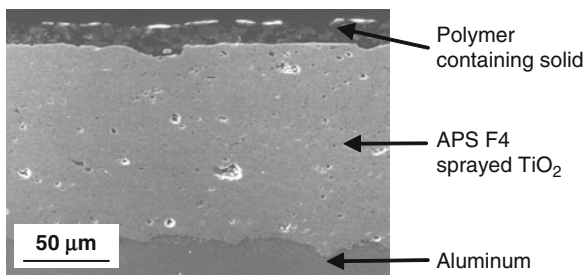
residual stresses in the TiO₂ coatings after the APS coating process depends strongly on the thermophysical properties of the substrate material and the heat and mass transfer during the deposition of the coating. With increasing

thermal expansion coefficient of the substrate material decreasing tensile residual stresses can be measured in the TiO₂ coating. The tribological investigation of APS sprayed TiO₂ coatings using a F1 torch and a high

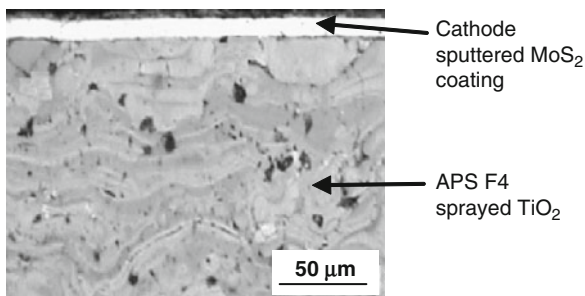
percentage of hydrogen indicate very interesting tribological properties concerning friction and wear coefficient. By means of XRD and WDS investigations substoichiometric TiO_x phases in the range of Ti_4O_7 and Ti_7O_{13} could be detected in these coatings.

Manufacturing and Evaluation of Composite Coatings with Thermally Sprayed TiO_x Primary Layers on Light Metal Substrates

For a further improvement of the tribological properties under dry friction conditions, TiO_2 coatings were deposited with a APS F4 torch as primary layers on AlSi12 substrates, and two different secondary coatings with dry lubricating ability were subsequently applied. One part of the samples was coated with lubricant lacquer containing microscale PTFE particles as solid lubricant by pneumatic air spraying (Gadow et al. 1998) (Fig. 11). The other part of the samples was coated with MoS_2 by cathode sputtering (Fig. 12). The coating thickness of the lubricant lacquer was $30\text{ }\mu\text{m}$ and the coating thickness of the MoS_2 coating was $2\text{ }\mu\text{m}$. The lubricant lacquer can either be



Self-Lubricating Treatment of Light Alloys, Fig. 11 SEM of a TiO_2 /lubricant lacquer composite coating on aluminum



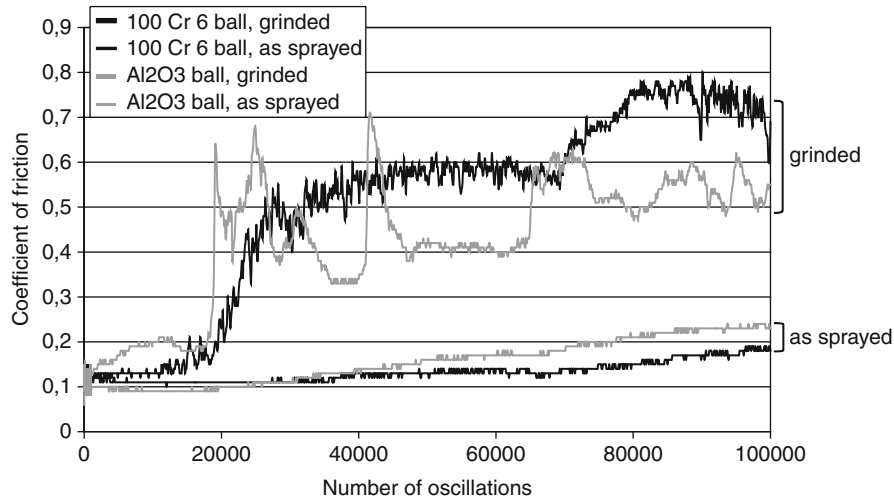
Self-Lubricating Treatment of Light Alloys, Fig. 12 Reflected light microscope shot of a TiO_2 /cathode sputtered MoS_2 composite coating

applied on the TiO_2 coating in the state “as sprayed” or the TiO_2 coating is polished prior to application of the lacquer. For the TiO_2 /cathode sputtered MoS_2 composite coating on the other side a polishing step of the TiO_2 coating is mandatory. A special sputtering process was applied using metallic interlayers between the individual MoS_2 solid lubricant layers to avoid a columnar structure growth of the MoS_2 layer (Nordbakke et al. 1998), since the lifetime of MoS_2 layers having columnar structures is quite limited.

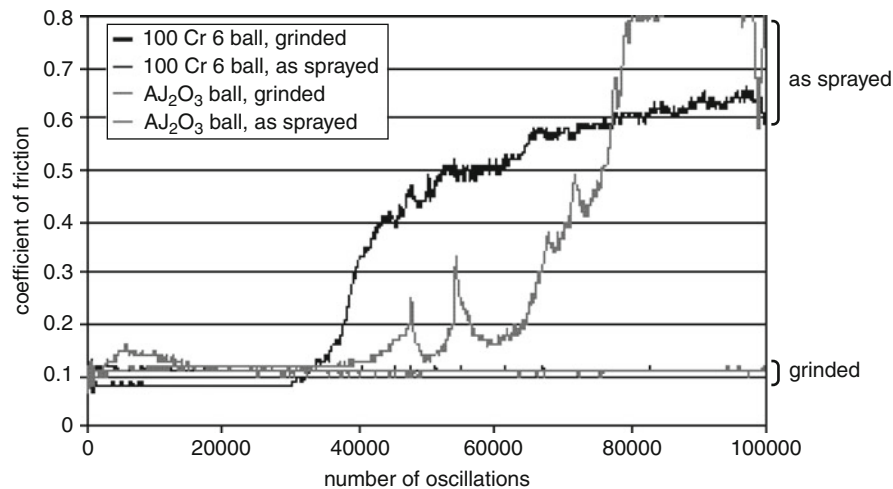
The tribological evaluation of the multilayer coatings was performed under dry friction conditions in the oscillating sliding mode as well. The same experimental parameters were chosen as for the other two test series, and Al_2O_3 as well as 100Cr6 balls (diameter 5 mm) were used as counterparts. The results of the tribological measurements are shown in Figs. 13 and 14. The TiO_2 /lubricant lacquer composite coating shows very low friction coefficients in the range of 0.1–0.2, even after 100,000 oscillations if the surface of the APS sprayed TiO_2 primary layer is not polished prior to the application of the lubricant lacquer (Fig. 13). If the TiO_2 layer is polished instead, the friction coefficient increases quite significantly already after 20,000 oscillations. This is due to the fact that a larger amount of lubricant lacquer can be deposited in the “as sprayed” surface, which is released constantly during operation. In addition, due to microabrasion, the most protruding asperities are leveled, thus lowering the specific pressure. For the polished samples, on the other hand, the applied lubricant lacquer layer is consumed fairly quickly and no “valleys” are available that could act as reservoirs. The scatter of the data for the coefficient of friction is an indication for three-body abrasive wear due to particles pullout forming wear debris that cannot be deposited in the “valleys,” as in the case of as sprayed coatings.

The tribological evaluation of the TiO_2 /sputtered MoS_2 composite shown in Fig. 14 demonstrates that, as expected, a deposition of the MoS_2 solid lubricant on the TiO_2 surface in the state “as sprayed” is not successful inasmuch as no running-in is possible and severe ploughing of the counterbodies and particles cracking due to high specific loads at the asperities causes severe wear and the friction coefficient increases quickly. For the polished samples, excellent low friction coefficients in the range of 0.1 are obtained even after 100,000 oscillations.

The tribological evaluation of the TiO_2 /lubricant lacquer composite coatings shows a lower friction coefficient if the thermally sprayed basic layer is used in the state “as sprayed” in comparison to using polished or machined surfaces. Since no post-treatment of the thermally sprayed layer is necessary, these combined coatings are very cost



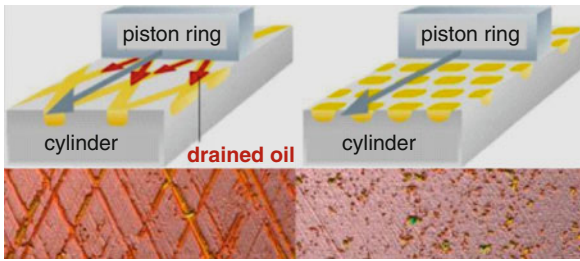
Self-Lubricating Treatment of Light Alloys, Fig. 13 Tribological behavior of TiO_2 /lubricant lacquer composite coating against 100Cr6 and Al_2O_3 balls under dry sliding conditions; surface roughness as sprayed $R_a = 4.0 \mu\text{m}$, $R_z = 24.9 \mu\text{m}$, surface roughness polished $R_a = 0.08 \mu\text{m}$, $R_z = 1.6 \mu\text{m}$



Self-Lubricating Treatment of Light Alloys, Fig. 14 Tribological behavior of TiO_2 /sputtered MoS_2 composite coating against 100Cr6 and Al_2O_3 balls under dry sliding conditions; surface roughness as sprayed $R_a = 3.0 \mu\text{m}$, $R_z = 18.8 \mu\text{m}$, surface roughness polished $R_a = 0.2 \mu\text{m}$, $R_z = 2.9 \mu\text{m}$

effective and functional TiO_2 -based multilayer systems. For the TiO_2 /sputtered MoS_2 composite coating a polishing step of the TiO_2 layer is necessary in order to exhibit low friction coefficients. These TiO_2 -based multilayer systems show an excellent long-term performance in contact with 100Cr6 and Al_2O_3 under dry sliding conditions. On the

other hand, the required post-treatment of the thermally sprayed layer as well as the application of the thin MoS_2 layer in a vacuum sputtering process are quite cost intensive. Thus, there is only a limited market for the application of these composite coatings in high value adding systems.



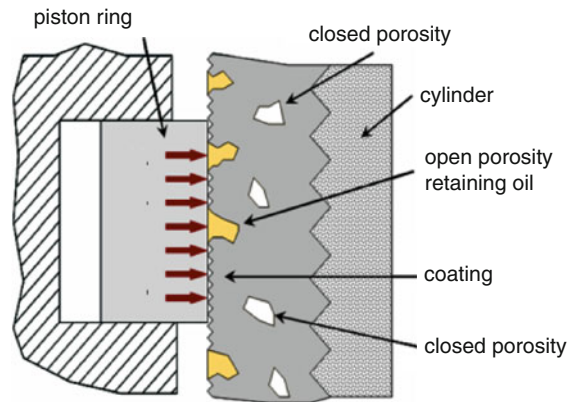
Self-Lubricating Treatment of Light Alloys, Fig. 15
Conventional honing grooves (left) vs. intrinsic porosity of thermal spray coatings (right)

Key Application

Cylinder Liner Coatings

Surface treatment of cylinder liners has been done for the last two decades in series production of internal combustion engines. The increased deployment of light metal alloys for engine crank cases made special surface treatments of the cylinder liners necessary to overcome the deficient wear resistance of the used materials. The most common methods of light metal alloy design are Nikasil® and Lokasil®, but thermal spray techniques like Rotaplasma®, plasma transferred wire arc (PTWA), high velocity oxy-fuel with wire (HVOF-w), and twin wire arc spray (TWAS) are gaining importance in cylinder liner treatment as an increasing number of passenger vehicles already feature thermally sprayed cylinder coatings in series production (e.g., engines from VW, Mercedes-Benz & AMG, Ford) (Barbezat et al. 1998; Barbezat 2005; Cook et al. 2003; Bobzin et al. 2008). Usually these coating processes only use Fe-based coating materials for effective protection of the light metal alloy crankcases against wear, but these coatings usually do not feature the high corrosion resistance much for engine operation with environmentally friendly biofuels. Other disadvantages of the established thermal spray techniques for cylinder coating are the dimensional constraints imposed by the necessity of an inside coating, leading to miniaturized spray equipment together with a small diameter of the cylinder, about 60 mm (Ciaravolo et al. 2006).

Another advantage of thermally sprayed coatings for cylinder walls is their microstructure. The intrinsic porosity of the lamellar buildup becomes open pores on the surface, creating separated voids for oil retention. These discrete micro-cavities act as pressure chambers that keep the lubricant in place more efficiently than the conventional interconnected honing grooves where the lubricant can drain from the load area through the grooves, leaving



Self-Lubricating Treatment of Light Alloys, Fig. 16 Porosity distribution in a cylinder coating

the piston ring in contact with the cylinder wall, (Schmid 2006) (Fig. 15).

Additionally, as the intrinsic porosity can be found evenly distributed throughout the coating thickness, the amount of pores and therefore the oil retention volume stays almost constant, even after excessive wear, as opposed to the honing grooves that get worn off during engine operation (Fig. 16).

Substrate Materials

The substrate materials for automotive applications are mainly Al- and Si-based. AlSi cylinder liners with typical passenger car dimensions as well as gray cast liners for truck engines can be coated. For different studies, cylinder liners and inline four-cylinder crankcases have been coated (Fig. 17, Table 3).

The Al liners were used for cylinder liner segment tests on a ball-on-disk tribometer as well as for piston ring segment tests. The Fe liners were used for tribometer and piston ring on cylinder segment tests and for fired engine tests. The four-cylinder crank cases were tested in fired and trailed engine test rigs.

Coating Materials

In order to show the diversity and the unrestricted material choice of the presented cylinder coating setup, a variety of materials have been investigated for their aptitude as cylinder liner coatings, including metal alloys, cermets, and ceramics (Table 4). The tested coating materials also include nanoscale components either processed by HVSEFS or by HVOF in agglomerated powders.

The Fe-alloy powder Durum NaNO₃® has been derived from a wire material already used for crank case coating by TWAS and features in-situ formation of



Self-Lubricating Treatment of Light Alloys, Fig. 17 Coated and tested cylinder liners and engines: gray cast iron liner (left), 2,000 cm engine (middle), 600 cm race engine (right)

Self-Lubricating Treatment of Light Alloys, Table 3 Investigated coating materials

Probe	Material	Dimensions (mm)
Al liner	AlSi9	Ø 82.5 × 120
Fe liner	Gray cast iron	Ø 131.0 × 250
4-cyl. engine	AlSi17	Ø 82.5 × 120
4-cyl. engine	AlSi17	Ø 67.0 × 50

Self-Lubricating Treatment of Light Alloys, Table 4 Investigated coating materials

Composition	Brand name	Size
Fe-alloy	DurumNano3	25 µm
FeCrMo	SulzerMetco Diamalloy 1008	36 µm
WC/Co	InframatInfralloy S7412	31 µm
Cr ₃ C ₂ /NiCr	GTV 80.81.1	28 µm
TiO ₂	Evonik P25	21 nm
TiC	H.C.Starck STD120	2 µm

nanoscale hard phases. The FeCrMo powder Diamalloy 1008 is similar to the material used for Rotaplasma[®] coatings but offers higher corrosion resistance due to its alloying components. The agglomerated WC/Co powder Infralloy S7412 is composed of submicron carbides with a 12% cobalt matrix. The 80.81.1 powder is composed of 75% microscale carbides and an 80/20 metal phase. The titanium oxide P25 and titanium carbide STD 120 were combined in a water-based dispersion with 30% solid matter in an 80/20 composition for the processing by means of HVSFS.

HVOF Coating Setup for Cylinder Liners/ Engines

The coating setup consists in a GTV TopGun[®] HVOF spray gun handled by a six-axis industry robot and a rotary table for the component handling (Fig. 18).

For the HVOF process, a commercial spray setup with the GTV TopGun[®] spray gun and GTV powder feeders (GTV PF2/2) has been used. Since the HVSFS process works with suspensions, a modified GTV TopGun[®] spray gun together with an IFKB Stuttgart designed suspension feeding equipment has been employed (Fig. 19). The HVSFS process offers the advantage of direct processing of nanoscale particles, but also demands a special design of critical elements, e.g., the combustion chamber and injection nozzle of the spray gun.

Coating Analysis and Characterization

As-Sprayed Coating Properties

The coatings have been analyzed as sprayed-on cylinder liners (GCI and AlSi). All coating characterization has been carried out at IFKB at the University of Stuttgart. The investigation included micro-hardness measurements determined on polished cross sections of the samples using a micro-indenter from Fischer, calculating a HV_{0.1} micro hardness according to ISO 14577, and coating porosity measurement by digital image analyzis of cross section micrographs (Table 5).

In addition, the as-sprayed surface roughness has been determined using a Mahr Perthometer for tactile surface scanning (Table 6).

As expected, the coatings show a higher hardness compared with state-of-the-art cylinder walls like CGI or AlSi. The cermet coatings (Cr₃C₂/NiCr and WC/Co) feature a very high hardness because of their carbide

hard phase. Due to these hardness values, a higher wear resistance of the coatings can be expected. The porosity values of the coatings show that the HVOF/HVSFS process actually produces dense coatings with a homogeneous phase distribution (Fig. 20).

Honed Coatings

All coatings have been smooth honed with very shallow honing grooves in order to have an even plane with the only structure on the surface originating from the open porosity of the coating. This open porosity is adequate as the only oil retention volume and is represented by the R_{vk} values of the surface measurement (Table 7).

The WC/Co-coating shows very low roughness and extremely low R_{pv} and R_{vk} values, resulting in a very smooth, mirror-like surface. Despite the smooth surface,



Self-Lubricating Treatment of Light Alloys, Fig. 18 Internal cylinder coating setup for crankcases at the IFKB

the WC/Co coating still holds enough open pores to guarantee sufficient oil retention.

Open Coating Porosity

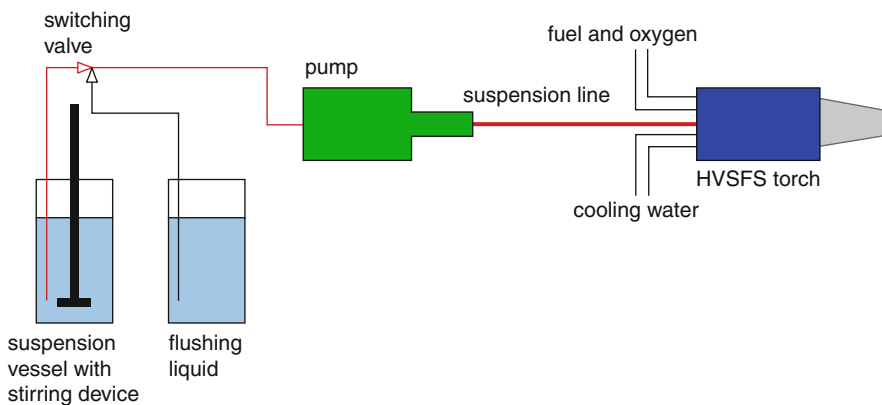
In addition to the tactile surface analysis, the open porosity on the honed coating surfaces has been investigated by means of high pressure mercury porosimetry (Porotec

Self-Lubricating Treatment of Light Alloys, Table 5 Micro hardness $HV_{0.1}$ and average coating porosity of the tested coatings

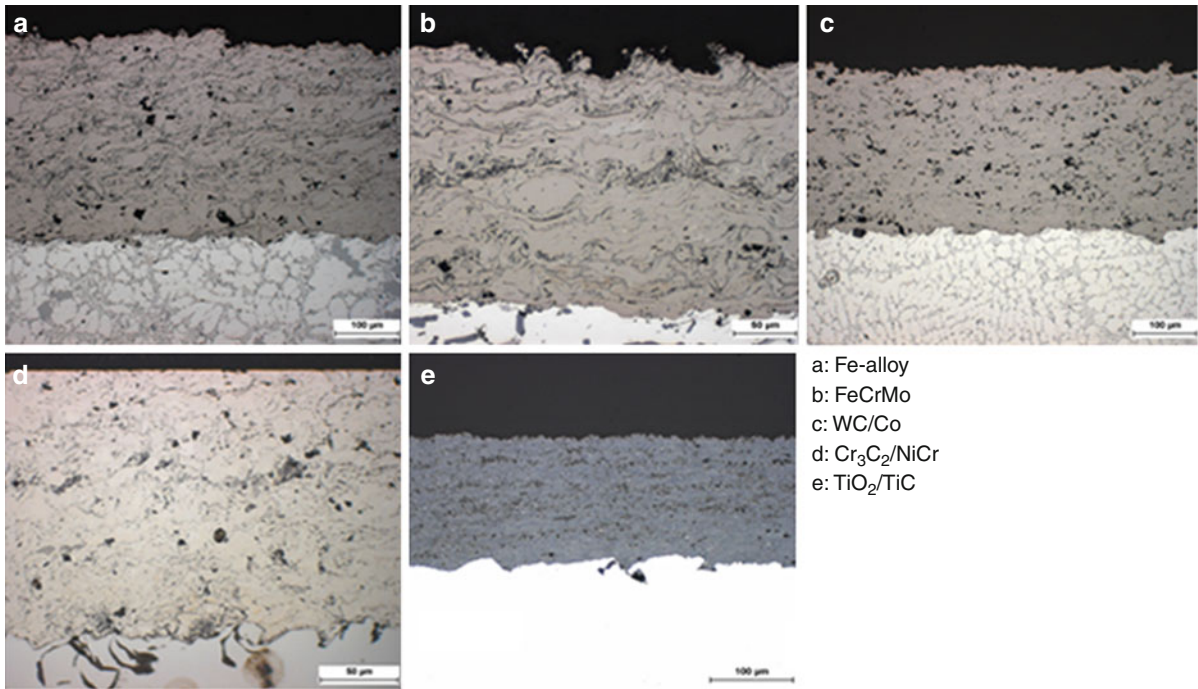
Coating	Hardness	Porosity (%)
Fe-alloy	810 $HV_{0.1}$	1.1
FeCrMo	710 $HV_{0.1}$	2.9
WC/Co	1,330 $HV_{0.1}$	0.1
$Cr_3C_2/NiCr$	1,000 $HV_{0.1}$	1.2
TiO_2/TiC	775 $HV_{0.1}$	1.4

Self-Lubricating Treatment of Light Alloys, Table 6 As-sprayed surface roughness of the coatings

Coating	R_a (μm)	R_z (μm)
Fe-alloy	7.94	51.15
FeCrMo	9.53	58.81
WC/Co	6.12	38.01
$Cr_3C_2/NiCr$	6.49	44.63
TiO_2/TiC	3.37	21.33



Self-Lubricating Treatment of Light Alloys, Fig. 19 Equipment for the HVSFS coating process, IFKB Stuttgart (Gadow et al. 2008)



Self-Lubricating Treatment of Light Alloys, Fig. 20 Cross section micrographs of the coatings

Self-Lubricating Treatment of Light Alloys, Table 7 Honed surface roughness of the coatings

Coating	R_a (μm)	R_z (μm)	R_{pk} (μm)	R_{vk} (μm)
Fe-alloy	0.20	4.99	0.05	1.06
FeCrMo	0.13	3.30	0.05	0.71
WC/Co	0.02	0.24	0.02	0.03
$\text{Cr}_3\text{C}_2/\text{NiCr}$	0.16	3.58	0.06	0.78
TiO_2/TiC	0.21	3.19	0.17	0.82

Pascal 140/440) in order to get detailed data on the pore size, pore volume, and pore distribution on the surface (Figs. 21, 22, Table 8). Therefore, coated cylinder wall segments were prepared and analyzed.

The measurements show that all coatings feature macro scale porosity as well as meso and micro pores that have not been detected by the cross section image analysis.

The pore size as well as their amount strongly depends on the coating material, which can also be stated from SEM images of the honed surfaces (Fig. 23). The aforementioned mirror-like WC/Co surface also contains pores, but in a small range between 0.2 and 1 μm in

diameter. The TiO_2/TiC coating shows rather shallow and cumulated pores whereas the $\text{Cr}_3\text{C}_2/\text{NiCr}$ coating features deep and somewhat sharp-edged pores with uneven distribution. The Fe-alloy and FeCrMo coatings both have flat and deep pores with sharp edges.

Test Results

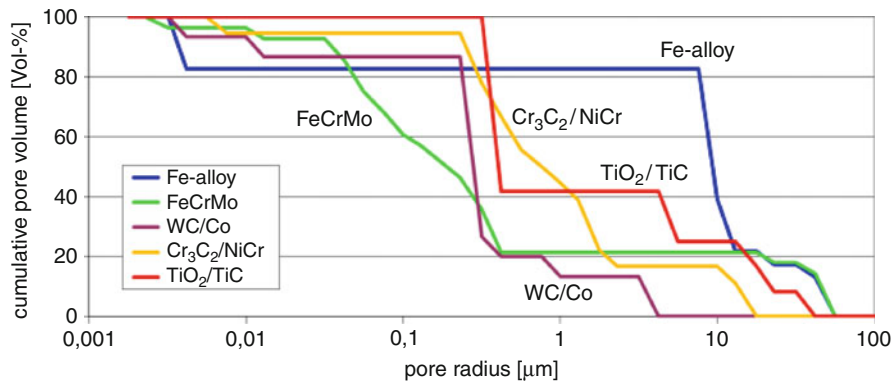
Lab Tests

The coated cylinder liners have been tested for wear resistance and friction coefficient on a tribometer with coated cylinder wall segments and compared with uncoated state-of-the-art cylinder wall segments.

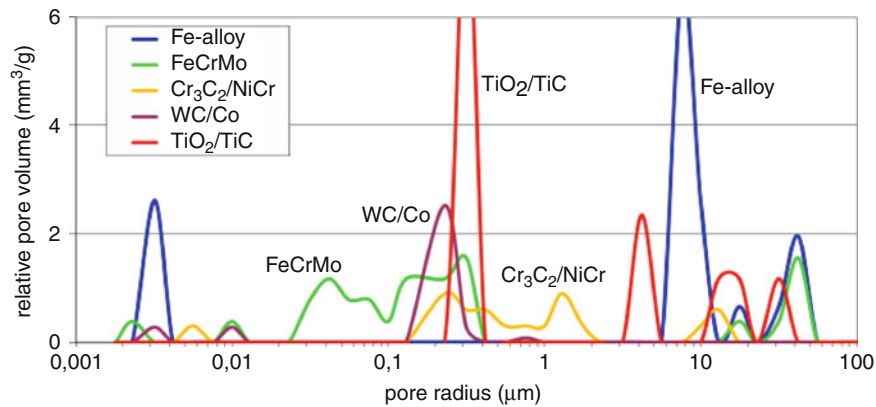
The first pin-on-disk test was conducted in non-lubricated conditions with an alumina ball as counterpart to accentuate the wear and load on the tested sample in order to have a time lapse effect for a simple identification of the wear (Fig. 24). The friction coefficient of the coatings has also been determined on the pin-on-disk tribometer (Table 9).

With the exception of TiO_2/TiC , all coatings showed a significant wear reduction compared with the standard cylinder wall materials. Especially the WC/Co coating showed extremely low wear.

The relatively high lubricated friction coefficient of the gray cast iron probe of 0.2 arises from the combination of



Self-Lubricating Treatment of Light Alloys, Fig. 21 Cumulative pore volume of the coatings



Self-Lubricating Treatment of Light Alloys, Fig. 22 Pore size distribution of the coatings

Self-Lubricating Treatment of Light Alloys, Table 8 Peak porosity volume and pore size range at porosity volume peak

Coating	Peak relative pore volume (mm³/g)	Pore size range at peak (μm)
Fe-alloy	5.0	5.0...11.0
FeCrMo	1.5	0.1...0.4
WC/Co	2.5	0.1...0.3
Cr ₃ C ₂ /NiCr	1.0	0.1...0.6
TiO ₂ /TiC	7.0	0.1...0.4

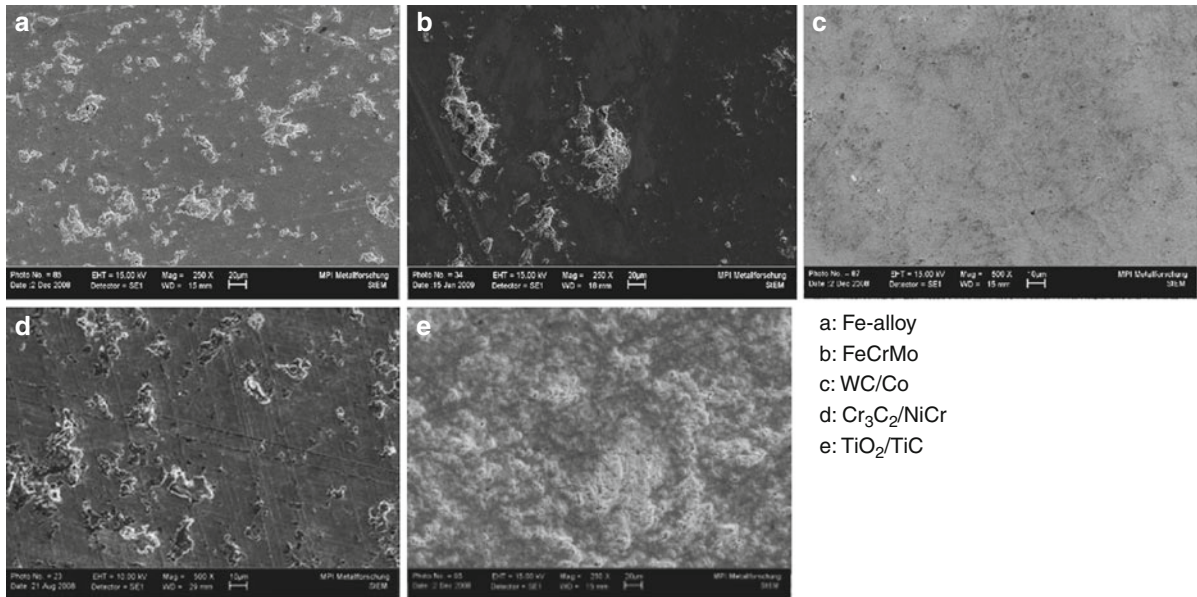
its rather deep honing grooves combined with the small contacting area of the 100Cr6 ball counterpart. The lower friction coefficients of the other probes show that the porosity and the shallow honing grooves of the AlSi

probe are better suited for holding the lubricant in areas of high load.

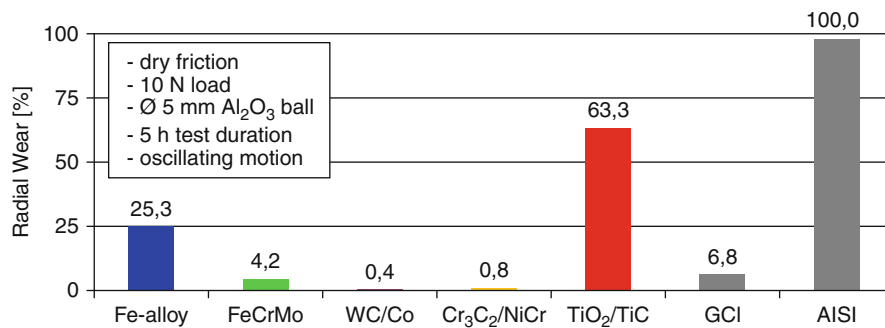
In addition to the somewhat theoretical pin-on-disk testing, a piston ring on cylinder segment wear test has been carried out in order to compare the wear of the coated cylinder walls to state-of-the-art materials like gray cast iron or a commercially available atmospheric plasma spray (APS) coating. This test has been conducted in a controlled environment with excess lubrication and under varying load and temperature states simulating different engine operation conditions. The wear measurement after the test showed a significant wear reduction for all coated probes compared to gray cast iron (Fig. 25). The TiO₂/TiC coating showed an outstanding wear resistance.

Engine Tests

For an explicit evaluation of the coatings, several engine tests have been conducted on trailed and fired engine test



Self-Lubricating Treatment of Light Alloys, Fig. 23 Exemplary SEM images of the honed coating surfaces with open porosity



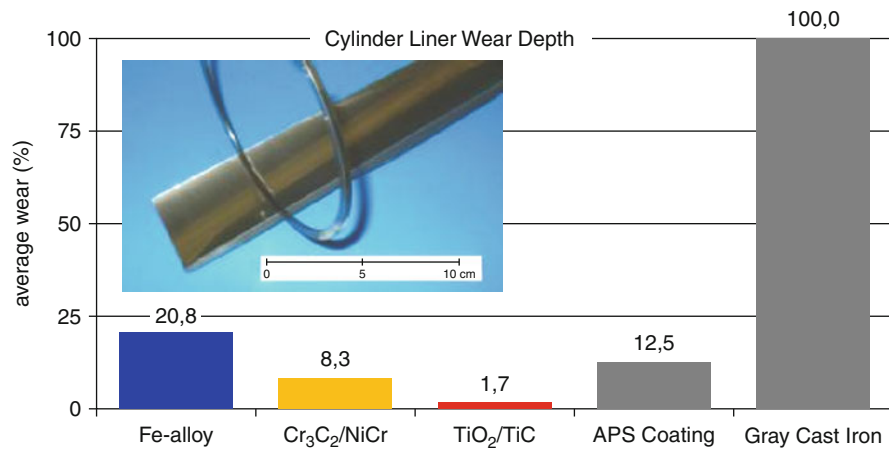
Self-Lubricating Treatment of Light Alloys, Fig. 24 Radial wear on the coatings after dry friction test

Self-Lubricating Treatment of Light Alloys, Table 9 Friction coefficient μ at lubricated conditions

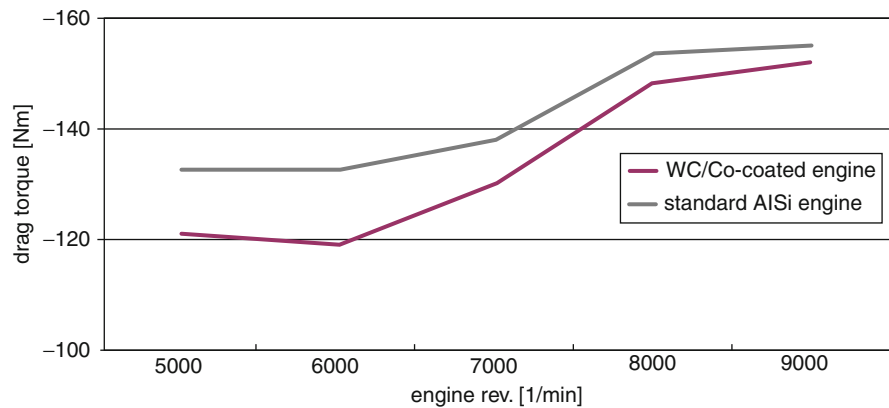
Coating	μ	Test conditions
Fe-alloy	0.10	– Excess lubrication – 10 N load – Ø 5 mm 100Cr6 ball – Oscillating motion
FeCrMo	0.11	
WC/Co	0.09	
Cr ₃ C ₂ /NiCr	0.10	
TiO ₂ /TiC	0.13	
Gray cast iron	0.20	
AlSi17	0.09	

rigs. A 600 cm race engine has been tested with a WC/Co coating and compared with the standard AlSi version on a trailed engine test bed to analyze the frictional losses (Fig. 26). The coated engine showed a reduction in drag torque of up to 10% (avg. 6%) depending on the revolutions per minute. This test result has been achieved without any further optimization of other engine components and therefore directly represents the friction reduction effect of the coating.

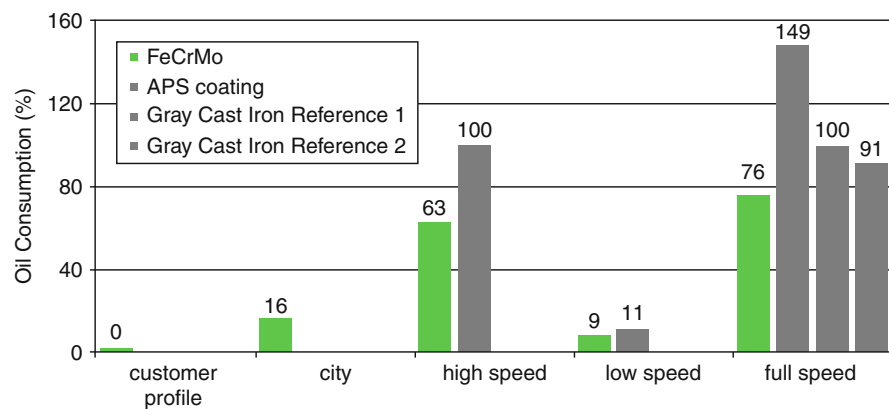
A second full engine test has been done with a FeCrMo-coated engine (2,000 cm) with the aim of analyzing the oil consumption (Fig. 27). The extensive



Self-Lubricating Treatment of Light Alloys, Fig. 25 Piston ring on cylinder segment wear (Volvo Technology Corp. Göteborg, Sweden, 2008)



Self-Lubricating Treatment of Light Alloys, Fig. 26 Drag torque measurement of coated and standard 600 cm engine (Formula Student, Team 0711-4, University of Stuttgart, 2010)



Self-Lubricating Treatment of Light Alloys, Fig. 27 Oil consumption of a FeCrMo-engine compared to reference engines (Audi AG Ingolstadt, 2008)

test cycle comprised several profiles with varying engine revolutions and load states to represent different driving conditions like urban driving or full speed phases. Compared with state-of-the-art technology like different gray cast iron configurations or commercial APS coating, the FeCrMo-coated engine showed considerably lower oil consumption throughout the test phases. Especially during a typical passenger car driving profile, the engine showed no detectable oil consumption and even 25% less consumption under full rotational speed at full load compared with the reference engine. This reduction of oil consumption also favorably impacts the emissions of the engine because less oil is being burned in the cylinder emitting less polluting substances.

When comparing the porosity measurements with the results from the tribometry and the engine tests, it can be stated that the top coatings all have a high amount of open pores in the same range of 0.2–1.0 μm diameter. This leads to the assumption that there might be a pore size range that is optimal for the frictional states that occur during engine operation. This optimum presumably supports best the hydrodynamic friction, even in areas with low relative motion between piston rings and cylinder wall (e.g., close to top dead center). Since the porosity is an intrinsic property of the coating, it renders additional surface structuring unnecessary reducing post treatment and time and cost.

The porosity structure depends on the employed material as well as on the coating process and the corresponding parameters, so an extensive investigation will be necessary to find the optimum pore size and distribution.

The tribological behavior of surfaces in internal combustion engines can be improved by metallic and ceramic protective coatings sprayed by HVOF and HVSFS. Tribological investigations show low friction coefficients and excellent wear resistance, in particular under difficult lubrication conditions. The intrinsic porosity of thermally sprayed coatings is able to hold oil very well and therefore form micro pressure chambers supporting the hydrodynamic state of friction.

Together with the high hardness of the tested coatings, the effect of the intrinsic porosity of HVOF-/HVSFS-sprayed cylinder wall coatings leads to an improved wear resistance, improved frictional conditions in the cylinder, and a reduction in oil consumption. Altogether, this technology has the potential of prolonging an engine's lifetime and reducing its emissions.

Acknowledgements

The contribution of Andrei Manzat and Andreas Rempp is kindly acknowledged.

Cross-References

- [Ceramic Coatings](#)
- [Ceramic Conversion of Light Alloys](#)
- [Contact of Layered Materials](#)
- [Engine Lubricants](#)
- [Environmentally Friendly Lubrication Issues](#)
- [Friction of Polymers](#)
- [Friction in Internal Combustion Engines](#)
- [Fuel Economy: Lubricant Factors](#)
- [PVD and CVD Coatings](#)
- [Self-Lubricating Hard/Ultra-Hard Coatings](#)

References

- S. Anderson, B. Collén, U. Kuylenstierna, A. Magnéli, Phase analysis studies on the titanium-oxygen system. *Acta Chem. Scand.* **11**, 1641–1652 (1967)
- G. Barbezat, Advanced thermal spray technology and coating for light-weight engine blocks for the automotive industry. *Surf. Coat. Technol.* **200**, 1990–1993 (2005)
- G. Barbezat, S. Keller, G. Wuest, Internal plasma spray process for cylinder bores in automotive industry, in *Proceedings of the 15th International Thermal Spray Conference*, Nice, 1998
- K. Bobzin, F. Ernst, J. Zwick, T. Schläfer, D. Cook, K. Nassenstein, A. Schwenk, F. Schreiber, T. Wenz, G. Flores, M. Hahn, Coating bores of light metal engine blocks with a nanocomposite material using the plasma transferred wire arc. *J. Thermal Spray Technol.* **17**(3), 344–351 (2008)
- M. Brune, M. Gramlich, Reibung und Verschleiß, in *Moderne Beschichtungsverfahren*, (DGM Informationsgesellschaft, Frankfurt a. M., 1996), pp. 213–230, ISBN 3-88355-223-2
- G. Ciaravolo, E. Witzgall, G. Mosetti, Application and evaluation of cylinder bore coatings for high-performance spark-ignition aluminum engines. *VDI-Berichte* **1906**, 247–258 (2006)
- D. Cook, C. Verpoort, K. Kowalsky, R. Dicks, Thermal spray of cylinder bores with the Ford PTWA process, in *VDI-Berichte Zylinderlaufbahn, Hochleistungskolben, Pleuel* (2003), Issue 1764, pp. 151–158
- R. Gadow, A. Killinger, D. Scherer, R. Gadow, A. Killinger, D. Scherer, Keramik-Polymer Kombinationsschichten. *Mat.-wiss. u. Werkstofftechn* **29**, 292–299 (1998). ISSN 0933-5137
- R. Gadow, A. Killinger, J. Rauch, New results in high velocity suspension flame spraying (HVSFS). *Surf. Coat. Technol.* **202**(18), 4329–4336 (2008)
- M.N. Gardos, M.N. Gardos, The effect of anion vacancies in the tribological properties of Rutile ($\text{TiO}_2\text{-x}$). *Tribol. Trans.* **31**(4), 427–436 (1988)
- W. Kleber, *Einführung in die Kristallographie* (VEB Verlag Technik, Berlin, 1967)
- M.W. Nordbakke, F. Heutling, M. Meyer, O. Knotek, M.W. Nordbakke, F. Heutling, M. Meyer, O. Knotek, New aspects regarding sputter-depositing dense coatings, in particular a solid lubricant of practical interest. *Mat.-wiss. u. Werkstofftechn* **31**, 205–214 (1998). ISSN 0933-5137 (in German)
- A. Plagge, Festschmierstoffe in Kraftfahrzeug-Schmierstoffen. *Tribologie und Schmierungstechnik* **38**(2), 74–81 (1991)
- J. Schmid, Optimierte Honverfahren für Gusseisen-Laufflächen. *VDI-Berichte* **1906**, 217–235 (2006)
- R. Schneider, Festschmierstoffe-Grundlagen und Anwendungsrichtlinien. *Schmierungstechnik* **18**(9), 280–285 (1987)

M. Woydt, A. Skopp, I. Dörfel, K. Witke, M. Woydt, A. Skopp, I. Dörfel, K. Witke, *Wear engineering oxides/anti-wear oxides*. *Wear* **218**, 84–95 (1998)

Self-Mating Ceramic Applications in the Hip Joint

J. GERINGER¹, K. KIM^{2,3}

¹Center of Health Engineering-Ecole Nationale Supérieure des Mines, Saint-Etienne, France

²School of Aerospace and Mechanical Engineering, Korea Aerospace University, Hwajeon-dong, Deogyang-gu, Goyang, Gyeonggi-do, Republic of Korea

³Materials Science Department, Penn State University, University Park, PA, USA

Synonyms

Biotribology of ceramic-ceramic hip implant; Ceramic on ceramic (CoC) articulations for hip replacement; Implanted biomaterials; Lifetime of ceramic hip joints

Definition

Ceramic hip implants replace the hip joint: femoral head-acetabulum (pelvic bone). Of four main hip joint couples, MoM, MoP, CoP, and CoC (M: metal; P: polymer, C: ceramic), CoC has increasing potential as a surface replacement for the near future. Improvement of this implant involves biotribological tests, such as the use of a hip simulator, for achieving increased durability. Moreover, specifically developed simulation (in vitro) devices are needed to take all in vivo conditions (and weaknesses of CoC couples) into account.

Scientific Fundamentals

Background

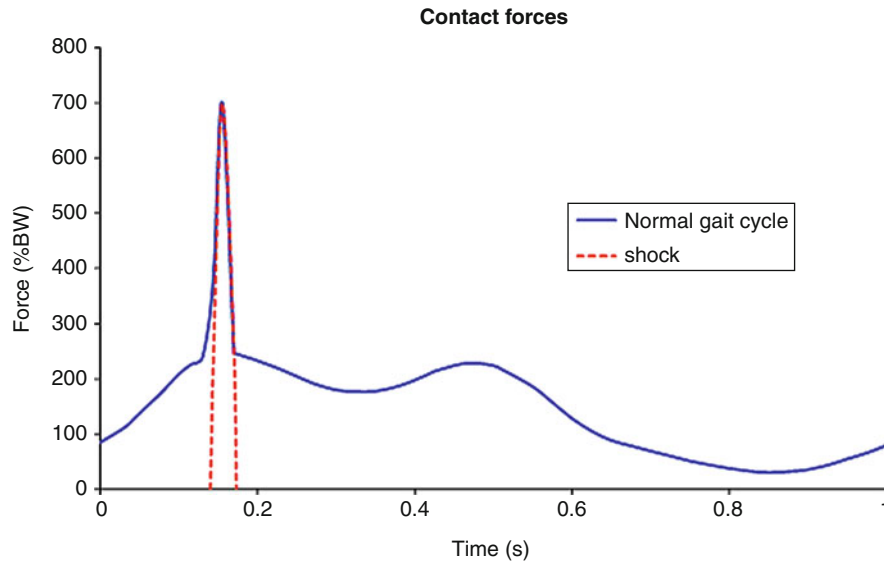
Since the 1970s, CoC couples have been implanted as hip joints. An orthopedic surgeon, Boutin, and his colleagues, focused their attention on this typical implant and implanted this material combination (Boutin 1972). CoC couples were constituted with alumina-alumina materials. It is worth noting the manufacturing process of this material. From the 1980s, the first surgical review of a CoC hip implant was published by Boutin et al. (1981). The fracture rate was on the same order of magnitude as the gold standard couple the MoP, implanted by Sir J. Charnley (1979) since the beginning of the 1960s.

After several iterations of surgical implantation and manufacturing processes, in the 1990s the CoC fracture rate dropped to approximately 0.004% for femoral heads (Schmalzried 2004). Although this fracture rate is low, it is worth noting that ceramic materials are brittle. Thus, when fractures of CoC hip implants occur, either the femoral head and/or the acetabular cup, the consequences are great and significantly impact the patient, who is submitted to a new surgical operation and health issues related to health organizations, government, insurances, and so on. The objective is avoiding fracture of this sort of material. From the 1980s onward, another ceramic material was investigated for replacing the deficient hip, zirconia. Until 2001–2002, the Prozyr® CoC was implanted as a promising material. However, more than 1,000 fractures occurred over the course of about a year, and this promising material was generally abandoned by the orthopedic community (Chevalier 2006). Advice and recommendations from health authorities all over the world was to avoid implantation of this kind of hip implant. From the beginning of twenty-first century, a composite material, alumina bulk and zirconia, has been investigated in order to take the best part of each material, and to benefit from synergisms between both materials.

Hip Biomechanics

This section highlights the weaknesses of CoC hip implants. Figure 1 describes the usual forces involved during gait (blue curve). It shows that seven times the body weight is possible as the maximum force (Uribe et al. 2011a).

This typical shape of forces vs. time during gait has to be taken into account for testing ceramic material. From the beginning of hip implant testing with hip simulators, only the blue signal, without the peak before 0.2 s, was applied between head and cup. Essner et al. (2005) showed that wear of the CoC couple is approximately 0.05 mm³/1 million cycles of gait when tested on a conventional hip simulator. This wear rate is approximately 1,000 times lower than the one for MoP bearings. From this point of view, one should conclude that CoC couples do not exhibit any degradation. It is worth noting that the mechanic characteristics of bioceramics are completely different than the ones of MoM or MoP couples; CoC is composed of brittle materials, unlike metal and polymer, and is under risk during impact loading. When the heel is not in contact with the ground, microseparation can occur. While head and cup are in close contact when the heel is on the ground, during the swing phase of the leg,



Self-Mating Ceramic Applications in the Hip Joint, Fig. 1 Contact forces involved during gait on the hip; BW body weight

the joint may disengage and microseparation is produced between head and cup (Tipper et al. 2002).

Physical and Mechanical Description

Table 1 describes a few physical and mechanical properties required from the ISO standard 6474 for alumina material and ISO 13356 for zirconia material. These properties come from manufactured specimens, and one might suggest that the materials properties of head and cup are the same. Thus, the manufacturing should involve no impact between the materials properties between head (alumina) and cup (alumina).

Alumina and ceramic materials are brittle. The elastic limit of typical ceramic materials lies between 0.4 and 0.6 GPa, and is close to 1 GPa for zirconia (Chevalier et al. 2009). Maximum stresses during gait are around 350 MPa, as suggested by modeling results (Uribe et al. 2011a). Thus, the ceramic material has to resist stresses involved during gait.

Another physical parameter is relevant: fracture toughness (Table 2). This physical parameter is related to growth of intrinsic defects in the material. The crack velocity related to a defect is usually drawn according to the stress intensity factor. Due to the brittle behavior of ceramic, when K_I , the stress intensity factor, reaches K_{IC} , the material breaks. Still, defects (quantity and shapes), even if they are rare in modern ceramics (less than 0.5%), are the weakest properties of ceramic material. Therefore, current activities focus on improving fracture toughness.

Self-Mating Ceramic Applications in the Hip Joint, Table 1

Physical and mechanical properties of alumina and 3Y TZP

	Average size of grains (μm)	Density ($\text{g}\cdot\text{cm}^{-3}$)	Young's modulus (GPa)	Elastic limit (MPa)
Head/alumina	1.49	3.97	402.0	400–600
Cup/alumina	1.50	3.97	404.5	400–600
Head/zirconia 3Y TZP ^a	0.175	6.08	210	1,000

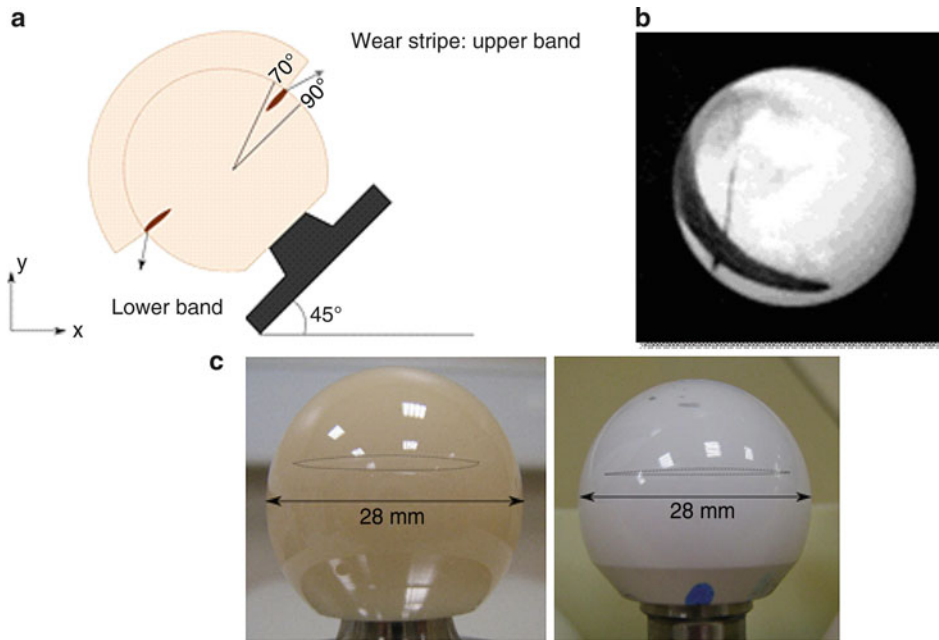
^a3Y TZP, 3% mol Ytria stabilized tetragonal zirconia polycrystals

Self-Mating Ceramic Applications in the Hip Joint, Table 2

K_{IC} and K_{IO} are, respectively, fracture toughness and fracture threshold; ZTA: zirconia-toughened alumina

	K_{IC} ($\text{MPa}\cdot\sqrt{\text{m}}$)	K_{IO} ($\text{MPa}\cdot\sqrt{\text{m}}$)
Alumina	4.2	2.4
3Y TZP	5.4	3.5
ZTA	6	5

Zirconia-toughened alumina (ZTA) is a composite material with bulk alumina and zirconia. This material exhibits the highest properties with regard to fracture toughness. As will be discussed later, it is likely the most



Self-Mating Ceramic Applications in the Hip Joint, Fig. 2 (a) Wear stripes on the ceramic head involved in in vitro shocks tests, upper and lower wear stripes. (b) Explants made of ceramics; implants' lifetime was less than 10 years (Magnissalis et al. 2001). (c) Femoral head after shocks tests: left image exhibits upper wear stripe on alumina head, right image exhibits ZTA head

promising new material. Degradation due to shock appears to be the “Achilles heel” of bioceramics in orthopedic implants, especially hip implants. The process could involve the following sequence: one step involves one shock due to microseparation. Thus, high contact forces are applied on the artificial joint, synonymous with high contact energy. Roughness is very low for ceramic material, due to localized defect or porosity in the bulk structure; and the contact stresses are locally very high, up to 350 MPa. From a shock, the defect grows until its size reaches the limit. The K_I value becomes close to the K_{IC} and finally the material breaks due crack propagation.

Key Applications

In Vitro Tests and Comparison with Explants

To reproduce the mechanisms of shocks that are the most damaging, a specific device should be developed to impose the shocks. The goal is to reproduce the red signal on Fig. 1 or the complete cycle. The entire blue signal is the most relevant. One might suggest that couples of other materials, MoM, MoP, and CoP, are involved in the assembly head-cup suffering from shock degradation. Indeed, metal and polymer, ultra-high molecular weight polyethylene (UHMWPE) are ductile, and shocks should promote, for

example, creep and fatigue degradation. Few studies are available in the literature about these couples. Additionally, MoM could suffer from degradation of its oxide layers and consequently might undergo corrosion. This phenomenon is close to erosion corrosion. For CoC couples, when shocks are applied during in vitro tests, experimental samples should be compared with explants. The comparison could be made in two ways, qualitatively and quantitatively. With regard to the former, Fig. 2a exhibits a scheme of degradation on the femoral head after shock tests. Figure 2b shows a black zone on an explanted head from Magnissalis et al. (2001). The wear zones are the same between in vivo and in vitro femoral heads. This point highlights that shock degradation is the relevant mechanism of ceramic degradation. Figure 2c shows macroscopic shock degradations on alumina and ZTA femoral heads. The black line shows the wear stripe, and it is evident that ZTA suffered lower degradation compared with alumina.

A quantitative comparison between explants and experimental samples to shock exposure provides actual data for identifying whether in vitro tests are consistent with ex vivo ones. Table 3 exhibits results related to wear rates of experimental samples and explants. From these results it is worth noting that wear rate of explants is

approximately the same as that related to the hip simulator (walking profile) and shock device. Considering shocks and hip walking simulator separately is not relevant to determining the actual wear rate. One may notice that the ZTA head, submitted to shock degradation, exhibits a wear rate that is 33% lower than that of an alumina head. Future ceramic hip joint research should focus on ZTA materials. Table 2 exhibits higher K_{IC} for the composite and Table 3 confirms this evolution with lower wear rate. ZTA and related developments around this material should provide interesting possibilities for future CoC implants. Actually, in vitro tests reproduce as close as possible ex vivo measurements for ceramic material. Moreover, shocks are a kind of initiator about the degradation of ceramic used for hip implants.

Role of Debris

The first challenge for studying debris, due to wear during gait, is its isolation from biological cells, proteins, and more generally tissues. Ceramic debris is small (grain size around

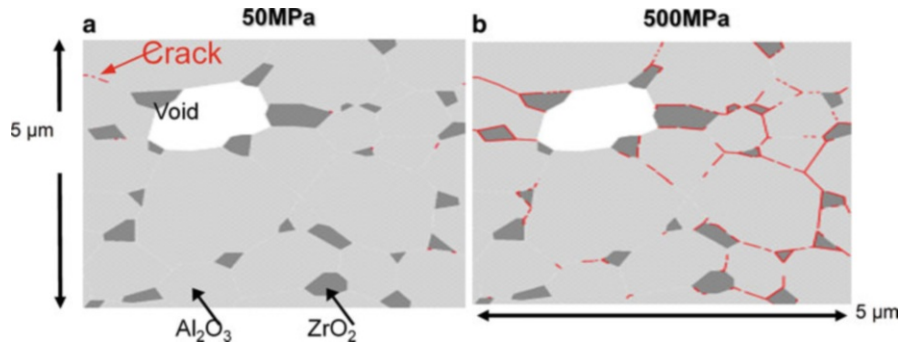
1 μm for alumina), and it can be even smaller than 1 μm due to intra-granular fracture. When debris is generated, interactions with biological media promote complex molecules and possible dissemination in tissue around the artificial joint. The inflammation of cells, osteoblasts, and/or fibro-nectin is not promoted with alumina-zirconia particles (Roualdes et al. 2010). It is worth noting that no deleterious and biological reactions were highlighted with alumina and zirconia particles. In order to confirm this point, the possible outlook should be the isolation of ceramic debris from actual hip implants. Afterwards, two bolts should be broken. The first one is related to the comparison between in vivo and/or ex vivo and in vitro; extracting debris from the biological medium in both cases. Thus, a statistical analysis should be performed to characterize the morphology of the debris. Secondly, the behavior of cells challenged with the collected debris should be very helpful in understanding the biological mechanisms of the reaction of cells due to debris.

Modeling

Modeling the degradation of bioceramics used in hip implants is challenging. The bulk structure is microscopic, even nanoscopic, and the mechanical specimen for example, of the femoral head is 28 mm of diameter. Figure 3 shows two-dimensional finite element modeling for investigating fracture and fatigue behavior of $\text{Al}_2\text{O}_3\text{-ZrO}_2$ ceramic composite at the microscopic level (i.e., ceramic grain scale). A composite structure was modeled with Abaqus® software. A simulation algorithm from model generation to analysis was detailed, enabling cyclic loading. A bilinear, time-independent cohesive zone law was implemented for describing fracture behavior of grain boundaries. The cohesive zone law allows elements in grain boundaries to be removed when they satisfy a predefined failure criterion. At the microscopic level, short and long cracks are

Self-Mating Ceramic Applications in the Hip Joint, Table 3
Wear rate of alumina and ZTA heads tested by shocks degra-dations; HWS: hip walking simulator; S: shocks; ex vivo: explants analysis

Reference	Wear rate, alumina head
(Essner et al. 2005)	0.02–0.08 mm ³ /million cycles/HWS
(Stewart et al. 2003)	0.2–1.84 mm ³ /million cycles/HWS + S
(Nevelos et al. 2001)	1 mm ³ /year/HWS + S
(Walter et al. 2004)	0.7 mm ³ /year/ex vivo
(Uribe et al. 2011b)	0.18 mm ³ /year/S
Reference	Wear rate, ZTA head
(Uribe et al. 2011b)	0.12 mm ³ /year/S



Self-Mating Ceramic Applications in the Hip Joint, Fig. 3 Alumina zirconia structure with void; red line represents cracks after applying normal load of (a) 50 MPa and (b) 500 MPa

observed among grain boundaries at the contact surface. The fracture phenomenon of a microstructure was reproduced by applying contact stresses to a model. This kind of modeling has to be improved in 3D and at higher scales, order of magnitude equal to 10 mm. This method should be a promising one for multi-scale modeling dedicated to hip implants. These future investigations will allow us to find the right design for this kind implant, head-cup-stem, and should be useful for better testing design. For example, when a Morse taper inside the femoral head is considered, some stresses are concentrated at the end of the cone. Thus, the chamfer has a key role in this kind of assembly and the multi-scale modeling (Kim et al. 2011) will be fruitful.

Cross-References

- [Self-Mating Metal Articulations in the Hip Joint](#)
- [Simulation of Physiological Conditions: An Overview](#)
- [Testing of Artificial Hip Joints](#)

References

- P. Boutin, Arthroplastie totale de hanche par prothèse en alumine frittée. *Rev. Chir. Orthop. Reparatrice Appar. Mot.* **58**, 229 (1972)
- P. Boutin et al., Le frottement alumine-alumine en chirurgie de la hanche. 1205 arthroplasties totales: Avril 1970 – juin 1980. *Rev. Chir. Orthop. Reparatrice Appar. Mot.* **67**, 279 (1981)
- J. Charnley, *Low Friction Arthroplasty of the Hip* (Springer, Berlin/New York, 1979)
- J. Chevalier, What future for zirconia as a biomaterial? *Biomaterials* **27**, 535 (2006)
- J. Chevalier et al., Ceramics for medical applications: a picture for the next 20 years. *J. Eur. Ceram. Soc.* **29**, 1245 (2009)
- A. Essner et al., Hip simulator wear comparison of metal-on-metal, ceramic-on-ceramic and crosslinked UHMWPE bearings. *Wear* **259**, 992 (2005)
- K. Kim et al., Two-dimensional finite element simulation of fracture and fatigue behaviours of alumina microstructures for hip prosthesis. *Proc. IMechE Part H J. Eng. Med.* **225**, 1 (2011)
- E.A. Magnissalis et al., Wear of retrieved ceramic THA components – four matched pairs retrieved after 5–13 years in service. *J. Biomed. Mater. Res.* **58**, 593 (2001)
- J.E. Nevelos et al., Wear of HIPed and non-HIPed alumina-alumina hip joints under standard and severe simulator testing conditions. *Biomaterials* **22**, 2191 (2001)
- O. Roualdes et al., In vitro and in vivo evaluation of an alumina–zirconia composite for arthroplasty applications. *Biomaterials* **31**, 2043 (2010)
- T.P. Schmalzried, How I choose a bearing Surface for My Patients. *J. Arthroplasty* **19**, 50 (2004)
- T.D. Stewart et al., Severe wear and fracture of zirconia heads against alumina inserts in hip simulator studies with microseparation. *J. Arthroplasty* **18**, 726 (2003)
- J. Tipper et al., Alumina-alumina artificial hip joints. Part ii: characterisation of the wear debris from in vitro hip joint simulations. *Biomaterials* **23**, 3441 (2002)
- J. Uribe et al., Degradation of alumina and zirconia-toughened alumina (ZTA) hip prostheses tested under microseparation conditions in a shock device. *Tribol. Int.* (2012) in press
- J. Uribe et al., Finite element modelling of shock-induced damages on ceramic hip prostheses. *ISRN Mater. Sci.* **1**, 1 (2011a)
- J. Uribe et al., Shock machine for the mechanical behaviour of hip prostheses: a description of performance capabilities. *Lubr. Sci.* **24**, 45 (2011b)
- W.L. Walter et al., Edge loading in third generation alumina ceramic-on-ceramic bearings: Stripe wear. *J. Arthroplasty* **19**, 402 (2004)

Self-Mating Metal Articulations in the Hip Joint

ALFONS FISCHER¹, SOPHIE WILLIAMS²

¹Werkstofftechnik, University of Duisburg-Essen, Duisburg, Germany

²Institute of Medical and Biological Engineering, School of Mechanical Engineering, University of Leeds, Leeds, UK

Synonyms

[Amorphous tribo-layers](#); [Beilby layer](#); [Fragmented layer](#); [Glaze layer](#); [Highly deformed layer](#); [Hip Joint](#); [Mechanically mixed material](#); [Nanocrystal layer](#); [Third body wear in a hip joint](#); [Transfer layer](#); [White-etching layer](#)

Definitions

A tribological system, or tribosystem, contains all substantial components contributing to tribological stresses, their properties, and changes under loading, as well as all characteristic mechanisms and magnitudes.

Tribomaterial – chemically and structurally altered material at the contact surfaces

THR – total hip replacement

HSR – hip surface replacement

Radial clearance – difference between the radii of head and cup

MoM – metal head on metal cup

CoC – ceramic head on ceramic cup

MoP – metal head on polymer cup

CoP – ceramic head on polymer cup

fcc – face centered cubic lattice structure

hcp – hexagonally closed packed lattice structure

SEM – scanning electron microscope

TEM – transmission electron microscope

EDS – energy dispersive X-ray spectroscopy

Scientific Fundamentals: The Tribosystem of Metal-On-Metal Hip Replacements

The artificial human hip replacement is a tribological system and should be examined using the system analysis methodology (Czichos 1978). In this methodology, the head is body 1 and the cup is the counterbody (Fig. 1) (Wimmer and Fischer 2007). The interfacial medium after surgery is the pseudo-synovia, which differs from the synovia in terms of its lubricating properties [Mazucco et al. 2002]. The environment is regulated by the human body. The operating variables load, relative speed, ambient temperature, and loading time bring about motion and work as an input to the system resulting in motion and work as output. The loss of the system can be defined in terms of energy (heat or sound) and material (wear debris). The loads and motions during daily activity determine the in- and output of the system and the moment of friction, as well as the generation and release of particulate wear debris; this and metal ions characterize the losses.

Materials for the Primary Articulating Surfaces of MoM Joints (Head, Cup)

Cobalt is known since 1735, however, the hexagonal closed packed lattice of pure cobalt hinders deformation and, therefore, makes it difficult to manufacture a wide variety of parts. Between 1907 and 1913, Elwood Haynes (Kokomo, Indiana, USA) invented cobalt-chromium alloys. These have a face-centered cubic lattice, are deformable and have a higher strength and a favorable cold working capability. In addition, chromium improves



Self-Mating Metal Articulations in the Hip Joint, Fig. 1 First generation MoM hip joint retrieved from an 80-year-old female patient after 8 years in situ. Head diameter is 35 mm (Büscher 2005)

the corrosion behavior. Haynes called these alloys “stellites,” because in daylight they took on a star-like appearance. The combination of mechanical and chemical properties was well understood for applications in chemical and mechanical engineering as well as in the mining industry, however, CoCr-alloys were not further developed. One reason for this was their toughness, which was lower compared with CrNi-steels available at that time. It wasn’t until 1937 that a CoCrMo-alloy called “Vitallium” was first used in dental implants to substitute gold, which was much more expensive. In 1938, CoCrMo-alloys were used in hip arthroplasty (Smith-Peterson 1939).

Today, many standardized CoCrMo alloys are applied in medicine, e.g., ISO5832-9, -6, -8, ASTM F-75, F-90, F1537, F562, and F563. These are based on Co, with 29% Cr and 6% Mo and additionally contain different amounts of other alloying elements (for example, C, Si, W, N, Ti, and B). They show tensile strengths between 900 and 1,920 MPa, while elongation to fracture values might range from 1% to 79%, depending on production process (cast, forged, sintered, hot-isostatically pressed, heat treated). The Youngs modulus of 210 GPa is similar to steel, but the endurance limit can be much higher, ranging from 190 to 900 MPa. A passive layer is generated within aqueous media, which is more stable than that of CrNi steels. CoCrMo-alloys are roughly distinguished by their C-content: HC- and LC-CoCrMo representing the high carbon alloys with more than 0.2% C and the low-carbon alloys with less than 0.04% C, respectively. While the latter consists only of face-centered cubic CoCrMoC solid solution with some small carbides (Fig. 2), the HC-alloys show additional eutectic Cr-carbides of Cr_{23}C_6 type precipitated at the grain boundaries (Fig. 3). The Cr-carbide size and distribution is highly dependent on the manufacturing route and sequence. In addition, some alloys may also show a complex mixture of Cr-carbides and MoSiCr-intermetallic phases (σ -phase) (Fig. 4), which have often been incorrectly designated as complex carbides.

Tribological Loading of Hip Joints

Daily activities of patients after total hip arthroplasty have been determined to consist of sitting (44.3% of the time), standing (24.5%), walking (10.2%), lying (5.8%), and stair climbing (0.4%), while resting periods between 2 and 30 s also contribute to a substantial part of daily activity [Morlock et al. 2001]. These may be detrimental because they cause a change from dynamic to static friction leading to higher frictional moments after rest (Nassut et al. 2003). The 10.2% walking fraction brings

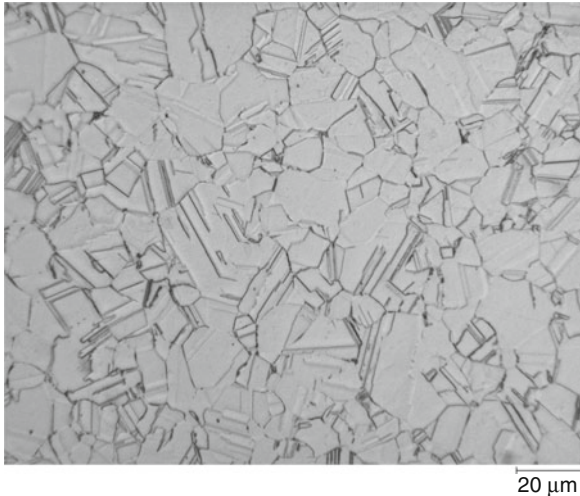
about one to three million gait cycles per year depending on the physiological age of the patient.

Force

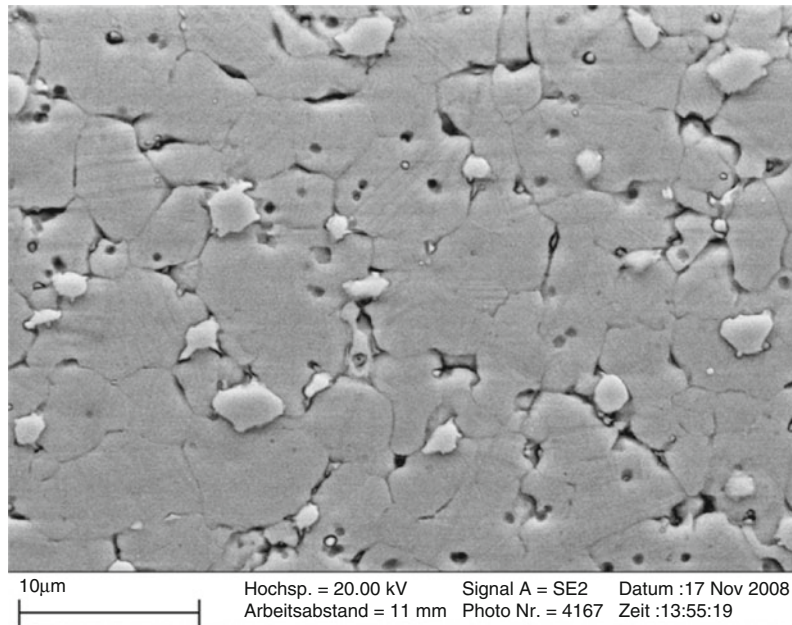
Bergmann et al. have measured contact force data during various activities using a telemetric transmission system

(Johnston et al. 2007). The average peak load at the hip was approximately 2.4 times the body weight (BW) during walking at a “normal” speed of 1.1 m/s. Ascending stairs, the joint contact force was 2.5 times BW and descending 2.6 times BW. The peak contact forces during all other common daily activities were comparably small; the exceptions were jogging (5.5 times BW) and stumbling (8 times BW).

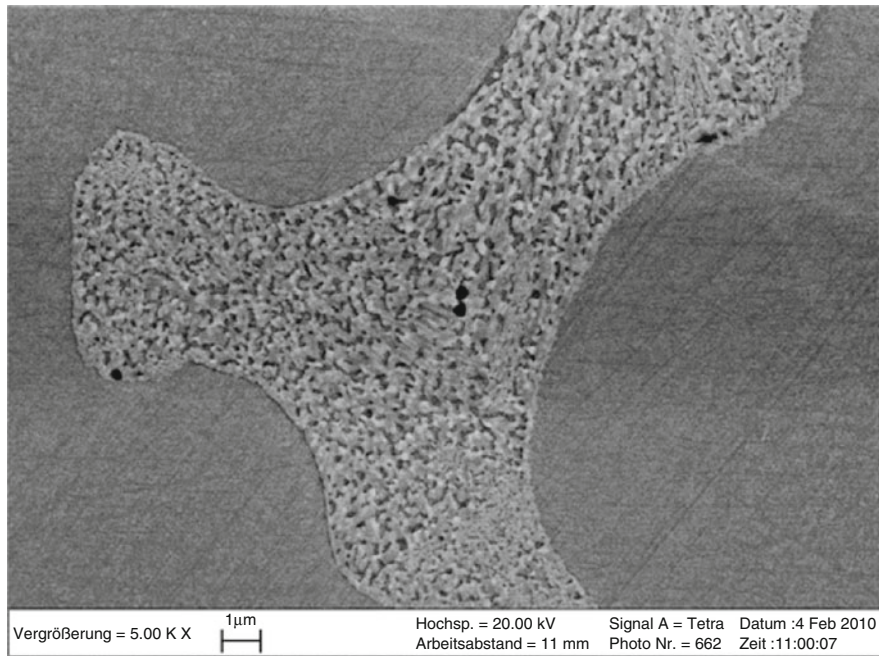
This normal force is transmitted via the real area of contact, which depends on the contact radii and clearance, the surface materials properties, and the topography. Since wear and friction arise from loading conditions and topography at the contacting interface, the properties of the surfaces in contact must be precisely considered. Technical surfaces are not smooth and exhibit varying degrees of roughness and waviness, which might range from the nanometer to the millimeter scale. The irregularities of the surface usually consist of broad based “hills” with very small angles of inclination, less than 15° from the base. The ratio of real to apparent area of contact changes constantly during motion and might range from 0.004 to 0.30. This depends directly on the statistical distribution of topographical irregularities, the acting shear, and normal contact forces, as well as the mechanical, physical, and chemical properties of the materials in contact. It should be noted that these surface properties may differ distinctly from the bulk properties. Thus, even



Self-Mating Metal Articulations in the Hip Joint, Fig. 2 Typical microstructure of a wrought LC-CoCrMo alloy with CoCrMo solid solution with some twins (light microscopy)



Self-Mating Metal Articulations in the Hip Joint, Fig. 3 Typical microstructure of a wrought HC-CoCrMo alloy with CoCrMo solid solution and eutectic $M_{23}C_6$ carbides precipitated at grain boundaries (SEM scanning electron microscopy)



Self-Mating Metal Articulations in the Hip Joint, Fig. 4 Mixture of $M_{23}C_6$ -carbides and MoSiCr-intermetallic phases precipitated at a grain boundary of a cast HC-CoCrMo alloy (SEM)

though the nominal Hertzian contact stresses might only reach values of about 50 MPa, the acting contact stresses within the real area of contact can be of several GPa. It is important to understand that the contact stress state changes from predominantly compressive to tensile at the surface and just below. Thus, every volume element experiences different alternating stresses and stress states. It is known that under a compressive stress metals tend to deform plastically at relatively small cyclic stresses. This is called cyclic creep or ratcheting and allows for the accumulation of large strains by small cyclic stresses.

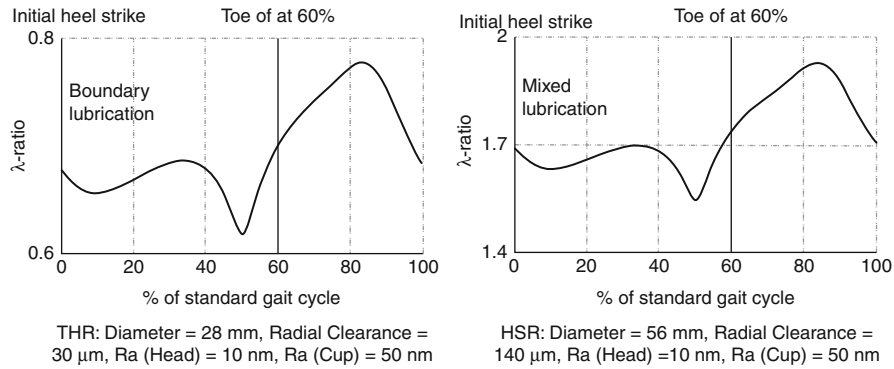
Ambient and Contact Temperatures

The ambient temperature that the bearings operate in is about 37°C, while the mean contact temperature has been measured in vivo and calculated to reach up to 51°C for metal-on-polymer couples. From in vitro tests, mean contact temperatures up to 74°C have been derived from MoP and up to 99°C for CoP and CoC couples. While in MoM hip replacements the mean temperature increases by 5–10°C above the ambient temperature, the flash temperatures, which only last for some microseconds, are calculated on the basis of in vitro data to reach maximum values of 60–80°C.

Relative Velocity, Entraining Velocity, and Lubrication

Relative motion combined with gait frequency and bearing geometry affect the relative velocity between contacting bodies during gait; relative motion involves all three angular degrees of freedom. At heel strike, the hip is in a flexed position and then extends throughout stance phase. During the time period from initial swing through mid-swing, the hip moves from extension into flexion and stays flexed until heel strike. For a typical person, the flexion-extension range of motion during stance spans from +30° (flexion) to –10° (extension) accompanied by a coronal arc movement of approximately 10° and transverse plane motion of approximately 15°. It should be noted that total hip replacement patients often show an abnormal gait pattern with decreased range of motion. Thus, the relative velocity range is between 0 and roughly 0.1 m/s during normal walking, while the velocity vector changes its direction between pure sliding and impact. Since both bodies move with different directions and speeds, the entraining velocity of the lubricant into the gap between head and cup, which is crucial for the state of lubrication, differs with the relative velocity.

The effectiveness of a lubricant film in a hip joint depends on many geometrical, physical, chemical, and



Self-Mating Metal Articulations in the Hip Joint, Fig. 5 Changes of the λ -value over a standard gait cycle of MoM joints. Notice! R_a has been measured at new retrievals and simulator specimens within the primary articulating areas. The shear thinning behavior of the pseudosynovia as well as the actual entraining velocities have been regarded

mechanical properties of all elements in the tribosystem. It can be defined by the parameter λ , and depends on factors including the diameter of the bearing, the radial clearance between the head and cup, the entraining velocity, the viscosity of the lubricant, the pressure across the articulating interface, and, last but not least, on the roughness of the surfaces. In the case of boundary lubrication ($\lambda < 1$), the lubricant adheres chemically to one of the surfaces and there is full contact between the solids, in contrast to hydrodynamic lubrication ($\lambda > 3$) where a total separation of the two bodies takes place. λ can be used to estimate the occurrence of different lubrication regimes (Hamrock and Dowson 1978). It should be mentioned that both load and relative velocity are not constants during one gait cycle. Equally, surface topography will vary over implant lifetime; it might decrease during run-in and increase due to wear or tribochemical reactions (Wimmer et al. 2003). In addition, the viscosity of the synovial fluid depends on its constituents, which have different effects on lubrication depending on the contacting materials and might change during the articulating action. Albumin, for example, tends to adsorb onto polymer surfaces, which changes the surface chemistry but has little influence on roughness. In MoM couples, proteins partly decompose, stick rigidly to the surfaces, and change the roughness. The decomposition of protein within a tribological contact is quite likely to be caused by a change in the environment, e.g., pH-value, temperature, mechanical stresses, or their superposition. Furthermore, it is known that synovial fluid is non-Newtonian and shows a distinct decrease in viscosity, over four orders of magnitude with increasing shear rate. Moreover, the tribological effect of the constituents of the pseudo-synovial fluid under boundary lubrication is

predominantly governed by proteins that are exposed or adsorbed onto the surface and only to a much lesser extent by those dissolved within the solvent. With MoM joints, λ is less than 1 and leads to predominantly boundary lubrication for hip replacement bearings, while for hip resurfacings, λ values between 1 and 2 are reached due to the much larger radii (Fig. 5). However, for worn and rougher surfaces this value decreases below 0.45 and hip resurfacings may run within the boundary regime as well.

Edge Loading

Under optimal biomechanical circumstances the real area of contact changes in terms of size and location on cup and head during one gait cycle. In particular, edge loading of the cup can occur in a variety of circumstances, for example, during when the head and cup become separated by some micrometers to millimeters. Under load, the repositioning brings about an edge contact at the superior or inferior rim of the cup at first, before the head slides into its final socket position. This edge contact leads to a distinctly smaller real area of contact and, therefore, increases the nominal contact pressure distinctly up to 700 MPa. Highly inclined and/or anteverted cups will also show this edge loading phenomenon. Thus, there is a distinct influence of edge loading on the wear behavior of hip joints, though the clinical factors influencing this are still under discussion. However, it should be incorporated into any in vitro testing in order to assure that the given combination is robust enough and forgives any biomechanically suboptimal conditions; this can be done using microseparation-type test regimes in hip simulators (Williams et al. 2004).

From the tribological loading one can conclude that contact forces, relative velocities, ambient and contact

temperatures, and their frequencies and durations are patient specific and highly variable; however, they will govern the initiation and progression of wear. In addition, many studies have revealed that after total hip arthroplasty, patients do not reach a so-called “normal” and age-specific walking pattern. Since most hip simulator wear tests are based on motion and force input data derived from “normal gait,” e.g., like in ISO or ASTM Hip Wear Testing standards, this must be considered when performing sound tribological analyses. Due to the nature of this tribosystem and the loading, one can assume that, besides a small amount of impact wear, the predominant type of wear is sliding wear. Thus, all four major wear mechanisms might appear. “Adhesion” (cold welding of surface spots by plastic deformation of surface asperities), “abrasion” (grooving of surfaces by plastic flow), “surface fatigue” (predominantly cyclic elastic deformation followed by crack initiation and propagation below the surface), and “tribochemical reactions” (chemical reaction between contact surfaces, interfacial media, and environment) are possible and it is important to investigate which of these are predominant and would consequently govern the wear behavior.

Low-Carbon Versus High-Carbon CoCrMo

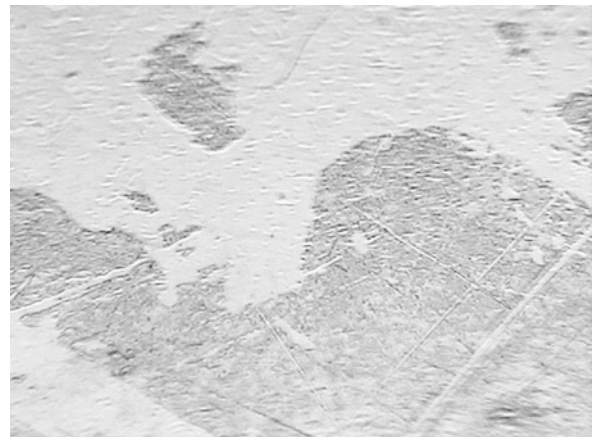
Low carbon content alloys produce significantly higher wear rates than high carbon content alloys in both simple configuration wear tests and hip-joint wear simulator tests. Hence, the pairing of low carbon cups with low carbon femoral heads is not recommended. High carbon/high carbon pairings show the lowest wear rates in hip-joint simulator tests (Dowson et al. 2004). The wear rates of cast and wrought CoCrMo alloys with and without various heat treatments have been compared and subject to debate. Dowson et al. reported no significant differences between the wear volumes of the wrought and cast high carbon CoCrMo materials (Dowson et al. 2004). However, the wrought material did exhibit a slight, non-significant advantage over the as-cast material. Heat treatments and hot isostatic pressing have been shown to have little effect on the wear rate of MoM hip prostheses. The effect of manufacturing route of MoM bearings on wear has been further considered under adverse wear conditions in hip simulator studies. Bowsher et al. investigated the wear generation of double-heat-treated and as-cast large-diameter MoM hip bearings using standard- and “severe”-gait simulations (Bowsher et al. 2006). High carbon MoM bearings (40 mm diameter) were manufactured and subjected to either hot isostatic pressing and solution annealing, or no heat treatment (as cast). No difference between the two groups under

both running-in and steady state conditions were observed, and the authors concluded that changes in alloy microstructure (due to manufacturing route) did not appear to influence the wear behavior of high carbon cast MoM articulations with similar chemical compositions.

From Wear Appearance to Acting Wear Mechanisms

The clinical linear wear rates for MoM hip replacements under steady state conditions range from 1 to about 8 μm per year. The wear process itself is represented by the generation of wear particles and ions and must be understood as a locally differing sequence of void initiation, accumulation, and propagation before wear particles, which are predominantly of some tens of nanometer in size, detach. Thus, according to a not-standardized but accepted definition, hip joints run in the range of so-called ultra-mild sliding wear that has a wear rate of below 10 nm/h.

MoM heads retrieved after 8–22 years typically exhibit reaction layers, indentations, and very few scratches (Fig. 6). These are characteristic for the combined action of the wear mechanisms’ “tribochemical reactions” and “surface fatigue”; there are few indications of “abrasion.” There are no signs of “adhesion,” which indicates that there must be something in the tribosystem that hinders direct metal-metal contact, even though these couples predominantly run under boundary lubrication. In addition, there must be a reason for the reported



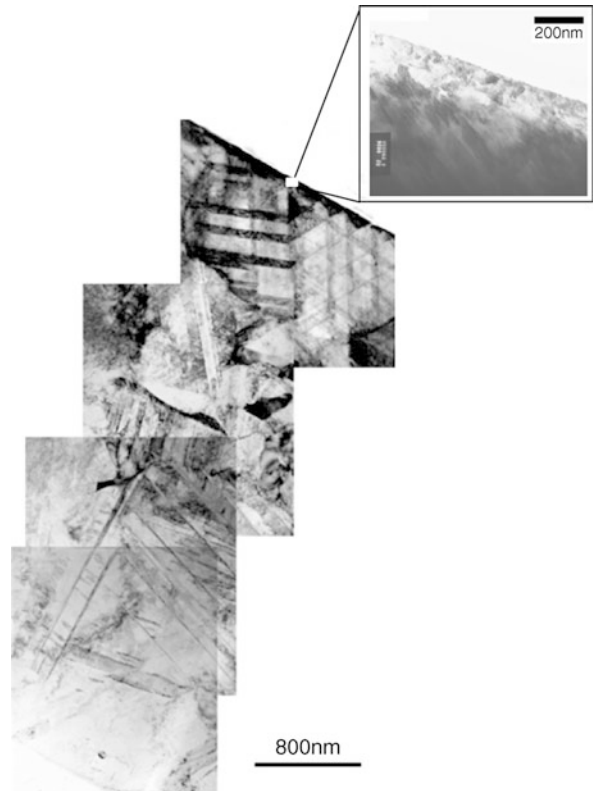
Self-Mating Metal Articulations in the Hip Joint, Fig. 6 Wear appearance of an articulating surface from a retrieval.

The darker areas represent tribochemical reaction layers while the lighter areas depict small indentations. Some grooves are visible as well (Büschler 2005) (SEM)

nanometer-small wear particles that are released into the body, even though the grain size of the base materials might range from 2 mm (cast prosthesis of the first generation) to 30 μm (today's clinically used cast and forged prostheses).

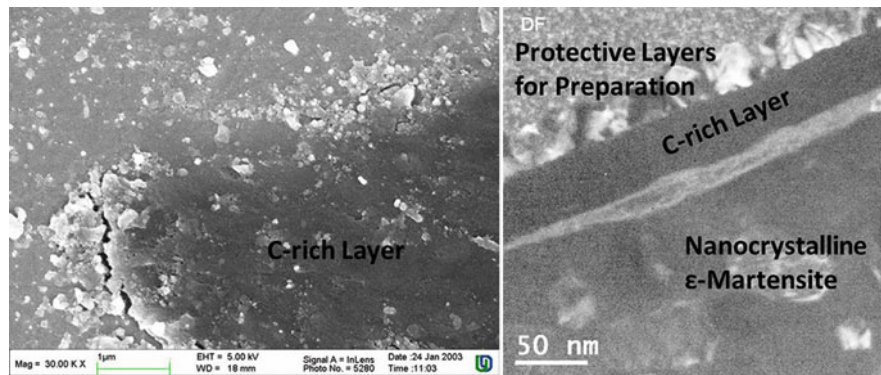
MoM bearings, as any other tribosystem, show a distinct run-in behavior in clinical applications, which leads to a larger number of particles being released during the first months or year compared with later periods of time. Thus, both contacting surfaces undergo a process of adjustment, before the head and cup are “settled” to conform under the peculiarities of the given biomechanical situation. With CoCrMo alloys this adjustment is brought about by an increasing density of lattice defects following the strain gradient towards the contact surface (Fig. 7). The base material is characterized by grains, which have a fcc lattice and already contain some lattice defects remaining from production. During tribological contact, further strain-induced twins, stacking faults, and a phase transformation into the hcp ϵ -martensite takes place closer to the contacting surfaces. With increasing shear strains towards the surface, this lathy martensite as well as stacking faults and twins can form rhombic domains, which are sheared, forming smaller ones or new nanometer-size grains. Due to the combined effect of all lattice defects in combination with the refinement of the effective grain size, the hardness increases as well, from 400 HV for the base material to about 1,000 HV near the surface. The adjustment of the micrometer-sized grains of bulk CoCrMo to nanometer-sized subsurface domains and grains is brought about by the cyclic stresses in combination with the compressive stress state. This leads to cyclic creep (ratcheting), allowing for the distinct strain gradient below the worn surfaces even though the frictional shear forces at such distance from the contact surface are relatively small, in the range of some tens of MPa. It is important to note that all micro- and nanocrystals generated by this mechanism have the same chemical composition as the bulk material.

Directly at the contact surfaces the situation changes. The grains remain 10–70 nm, but at depths between 0 and 300 nm from the surface a different mechanism prevails. Here, the differences in motion of head and cup have to be accommodated by a thin layer of solid material undergoing shear rates of 10^3 – 10^5 s^{-1} . Rigney and Karthikeyan (2010) showed by means of molecular dynamic simulations that at the interface of sliding bodies in contact, a friction-induced rotation of randomly arranged clusters of atoms or spontaneous generation of rotating nanocrystals takes place. This may cause material transfer, however, this type of material transfer would not lead to

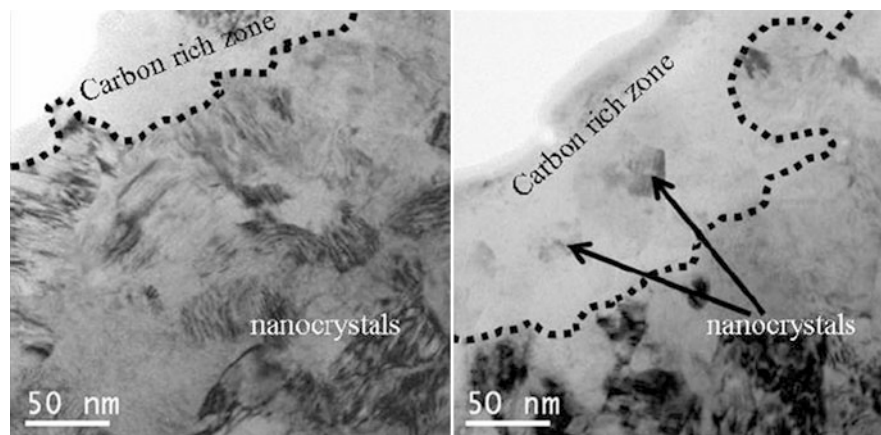


Self-Mating Metal Articulations in the Hip Joint, Fig. 7 Microstructural gradient from the microcrystalline base metal (bottom) to the nanocrystalline near-surface material (top). Notice that the deformation is solely brought about by plane slip with twins, stacking faults, and strain-induced ϵ -martensite. The uppermost layer in the magnified surface picture shows the tribomaterial (By Courtesy of Robin Pourzal, Duisburg, Germany and Mark Rainforth, Sheffield, UK) (TEM transmission electron microscopy)

the cold welding of hip replacement bearing surfaces; that would be attributed to “adhesion” and severe wear appearances or even stiction would be observed. In this instance, the transfer affects only a very small volume of material and does not distinctly hinder sliding, even though energy is dissipated and, therefore, friction forces apply. At first this mechanism will bring about a zone of rotating matter, which is also known for the plastic deformation of nanocrystalline metals under shear, and, in parallel, the interfacial medium between the body and counterbody is incorporated into this uppermost layer and consequently leads to a change in the chemical composition. This process is called mechanical mixing and does not generate a new alloy but a metallo-organic composite of organic



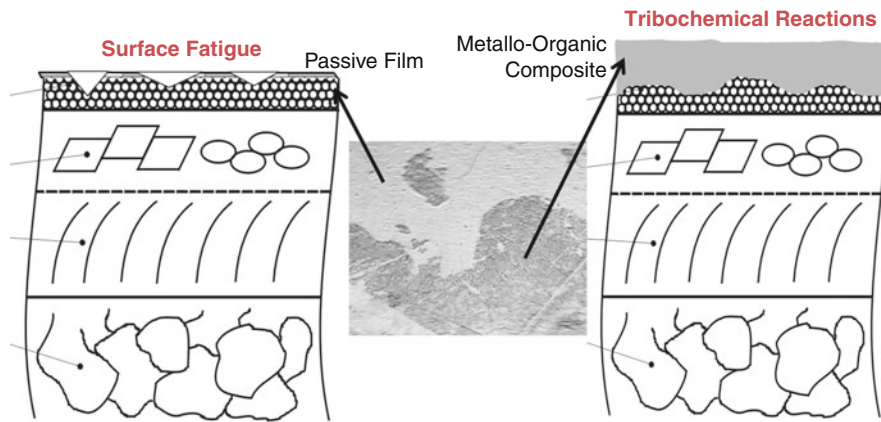
Self-Mating Metal Articulations in the Hip Joint, Fig. 8 SEM picture of a denatured protein film (*left*) as well as wear debris sticking rigidly onto a oxide layer from a retrieved cup (Büschler 2005) and TEM cross section (*right*) from a head of a CoCrMo hip simulator specimen beneath the primary articulating area showing the Cr-oxide layer between the worn metallic surface and the denatured protein designated as C-rich area according to energy-dispersive X-ray spectroscopy (EDS) measurements (By Courtesy of D. Raabe, R. Seeger, MPIE GmbH, Düsseldorf and Robin Pourzal, Duisburg, Germany)



Self-Mating Metal Articulations in the Hip Joint, Fig. 9 TEM cross section from a head of a CoCrMo hip simulator specimen within the primary articulating area depicting the tribomaterial. Carbon acc. to EDS from denatured proteins of the bovine serum is mixed into the contact surface by mechanical mixing together with the nanocrystals of the underlying shear fatigued metal (Courtesy of Robin Pourzal, Duisburg, Germany)

matter, with proteins from the pseudosynovial fluid and nanocrystals out of the uppermost metal. Thus, proteins might either stick rigidly onto an oxide layer on the worn surface (Fig. 8) or are incorporated into the uppermost surface volume (Fig. 9) (Wimmer et al. 2003). The latter is the nanostructured metallo-organic composite called tribomaterial, which is attributed to being generated by this submechanism of “tribochemical reactions.” During steady state wear, this nanostructured compound layer is worn away by “surface fatigue” brought about either by blunt asperities or by rotating nanosize wear debris.

The blunt asperities sliding over the counterface cause surface fatigue, which is stress-driven and characterized by predominantly cyclic elastic deformation. The repeated multiple indentations fatigue the material strain-driven, while at every contact the plastic deformation is most distinct. As this takes place either within the metallo-organic composite or the nanocrystalline ϵ -martensite, the wear particles of MoM couples resemble the size of the former nanocrystals, however, they tend to agglomerate so larger particles can be observed. Their size may range from 20 to 500 nm, while most of them are smaller



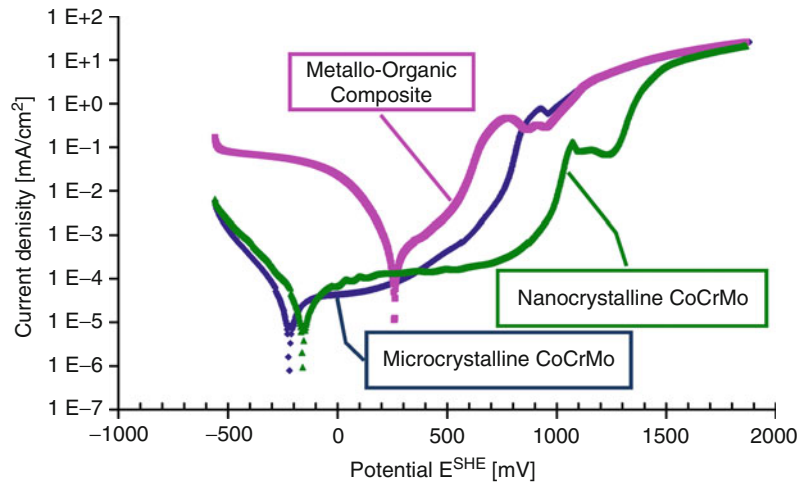
Self-Mating Metal Articulations in the Hip Joint, Fig. 10 Scheme of the two main wear mechanisms allowing for the ultra-mild sliding wear of MoM joints. Notice that this is a floating balance between two different contact situations depending on the individual contact characteristics

than 100 nm and have a globular shape. Sometimes needle-like particles are found as well, which are fractured ϵ -martensite lathes that have been torn off the surfaces in areas undergoing higher contact forces. Even though loose particles characterize this contact situation, it is totally different from three-body abrasion because, in a hip joint, these particles are generated inherently and are not added from the outside into the tribosystem. Whether particles generate scratches (abrasion) or indentations (surface fatigue) depends on their size and shape. The compact shaped nanosize particles from MoM hip replacement surfaces are very likely to rotate during contact, while agglomerates might also slide. Sliding particles generate grooves and wear away surface material in one cycle instead of by multiple contacts, and this would cause the wear rate being orders of magnitude higher and would not allow for ultra-mild sliding wear. The same is true for particles introduced from the outside of the immediate tribosystem (for example, by bone cement or bone). Therefore, only this specific combination and balance between the submechanisms of “tribochemical reactions” and “surface fatigue” taking place in the nm range allow for ultra-low wear rates observed in MoM bearings (Fig. 10).

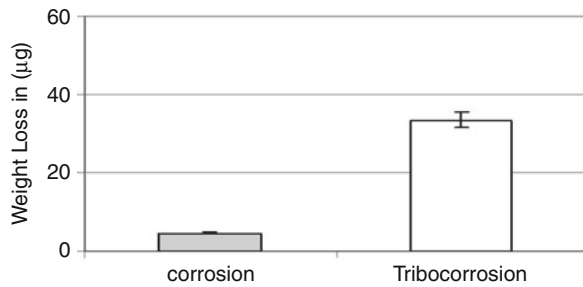
It cannot be concluded from these findings that MoM bearings are robust against suboptimal positioning as well as other sources of overloading, because the applied materials can adjust themselves to the acting stresses. It has been found that this metallo-organic tribomaterial is sensitive to loading. On surfaces that articulate under higher contact stresses as a consequence, for example,

suboptimal positioning, larger radial clearances or edge loading, the tribomaterial is missing. Thus, tribochemical reactions do not contribute substantially to such systems. As a result, mechanically dominated wear mechanisms such as “surface fatigue” prevail, causing increased removal of material and the wear rate increases by some orders of magnitude.

Nevertheless, even with a linear wear rate as small as $1 \mu\text{m}$ per year some 10^{12} particles are released into the body, which have a certain osteolytic potential (Ingham and Fisher 2005). This is accompanied by the release of metal ions due to steady depassivation/repassivation processes at the surface, during which bare metal surfaces are exposed to the synovia for a short period of time. Tribologically induced depassivation depends on the plastic deformation of the surface layers while the repassivation is hindered by proteins, which adsorb directly to fresh metal surfaces. Polarization curves of microcrystalline CoCrMo, nanocrystalline CoCrMo, and the metallo-organic composite are shown in Fig. 11. Obviously the micro- and nano-crystalline CoCrMo are quite similar, while the mechanically mixed metallo-organic composite appears nobler as to the higher corrosion potential, while in contrast the passive current density is higher as well. The latter might be related to the missing oxide layer between the metallo-organic composite and the underlying metal, which allows for a faster transport of ions towards the surface. The reasons for this shift of the corrosion potential are not yet clear. Still, the tribocorrosion taking place at the metallic areas of the contact surfaces leads to an increased release of ions by



Self-Mating Metal Articulations in the Hip Joint, Fig. 11 Polarization curves of micro- and nanocrystalline CoCrMo in comparison with the metallo-organic composite in new born calf serum with 30 mg/mL protein (By Courtesy of Michael Schymura, Duisburg, Germany and Mathew Mathew, Chicago IL, USA)



Self-Mating Metal Articulations in the Hip Joint, Fig. 12 Comparison of the weight loss of HC-CoCrMo by corrosion and tribocorrosion after 100 cycles of reversed sliding against a 28 mm ceramic head at 1 Hz and 16 N load (By Courtesy of Mathew Mathew and Markus Wimmer, Chicago IL, USA)

the corrosive interaction of a surface in a tribological contact by tribocorrosion (Fig. 12).

In order to achieve a sufficient coincidence between in vivo and in vitro results the characteristic acting wear mechanisms must match. Further, the laboratory tribosystem must be defined well as to its wear rate and the particle characteristics and allow for edge contact.

MoM bearings can show a low wear rate in vivo. However, due to the amount of released nanosize particles and ions there is a risk clinically of an adverse reaction to this or hypersensitivity-type reaction. This balance of advantages and disadvantages is true for all hip replacement material combinations.

Key Applications

The key applications lie in total metal-on-metal hip prostheses and hip replacements. The shown system approach as well as the understanding of acting mechanisms and submechanisms can be applied to any tribosystem. The combination of tribomaterial and underlying shear zone is crucial for so-called ultra-mild wear in any boundary and mixed lubricated metal-on-metal system, e.g., in automotive, tooling, mechanical, and biomedical engineering.

Cross-References

- ▶ [Boundary Lubricants](#)
- ▶ [Contact Temperature of a Moving Solid Surface](#)
- ▶ [Lubrication Modeling of Artificial Hip Joints](#)
- ▶ [Lubrication Regimes](#)
- ▶ [Sliding Wear](#)
- ▶ [Stochastic Contact Theories: Other Theories Based on the Greenwood-Williamson Model](#)
- ▶ [Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics](#)
- ▶ [Stresses in Contacting Materials](#)

References

- J.G. Bowsher, J. Nevelos, P.A. Williams, J.C. Shelton, 'Severe' wear challenge to 'as-cast' and 'double heat-treated' large-diameter metal-on-metal hip bearings. *Proc. Inst. Mech. Eng. Part H – J. Eng. Med.* **220**(2), 135–143 (2006)
- R. Büscher, *Gefügeumwandlung und Partikelbildung in künstlichen Metall/Metall-Hüftgelenken*, Ph.D. Thesis, University Duisburg-Essen, Duisburg (VDI Verlag, Düsseldorf, 2005), s.a. *Fortschr. Ber. VDI Z* 17(256)

- H. Czichos, *Tribology – A Systems Approach to the Science and Technology of Friction, Lubrication and Wear*. Tribology Series, vol. 1 (Elsevier, Amsterdam, 1978)
- D. Dowson, C. Hardaker, M. Flett, G.H. Isaac, A hip joint simulator study of the performance of metal-on-metal joints. Part I: the role of materials. *J. Arthroplasty* **19**, 118–123 (2004)
- B.J. Hamrock, D. Dowson, Elastohydrodynamic lubrication of elliptical contacts for materials of low elastic modulus I – fully flooded conjunction. *Trans. ASME J. Lubr. Technol.* **100**, 236–245 (1978)
- E. Ingham, J. Fisher, The role of macrophages in osteolysis of total hip replacement. *Biomaterials* **26**, 1271–1286 (2005)
- J.D. Johnston, P.C. Noble, D.E. Hurwitz, T.P. Andriacchi, Biomechanics of the hip, in *The Adult Hip*, ed. by J.J. Callaghan, A.G. Rosenberg, H.E. Rubash, vol. I, 2nd edn. (Lippincott Williams & Wilkins, Philadelphia, 2007), pp. 81–90
- D. Mazucco, G. McKinley, R.D. Scott, M. Spector, Rheology of joint fluid in total knee arthroplasty patients. *J. Orthop. Res.* **20**, 1157–1163 (2002)
- M. Morlock, E. Schneider, A. Bluhm, M. Vollmer, G. Bergmann, V. Muller et al., Duration and frequency of every-day activities in total hip patients. *J. Biomech.* **34**(7), 873–881 (2001)
- R. Nassut, M.A. Wimmer, E. Schneider, M. Morlock, The influence of resting periods on friction in the artificial hip. *Clin. Orthop. Relat. Res.* **47**, 127–138 (2003)
- D. Rigney, S. Karthikeyan, The evolution of tribomaterial during sliding: a brief introduction. *Tribol. Lett.* **39**(1), 3–7 (2010)
- N.N. Smith-Peterson, Arthroplasty of the hip: a new method. *J. Bone Joint Surg.* **21**, 269 (1939)
- S. Williams, T.D. Stewart, E. Ingham, M.H. Stone, J. Fisher, Metal-on-metal bearing wear with different swing phase loads. *J. Biomed. Mater. Res. Appl. Biomater.* **70B**, 233–239 (2004)
- M.A. Wimmer, A. Fischer, Tribology, in *The Adult Hip*, ed. by J.J. Callaghan, A.G. Rosenberg, H.E. Rubash, vol. I, 2nd edn. (Lippincott Williams & Wilkins, Philadelphia, 2007), pp. 215–226
- M.A. Wimmer, C. Sprecher, R. Hauert, G. Täger, A. Fischer, Tribochemical reaction on metal-on-metal hip joint bearings – a comparison between in-vitro and in-vivo results. *Wear* **255**, 1007–1014 (2003)

Self-Pressurized Hydrostatic Bearing

- [Biphasic Lubrication](#)

Self-Sintered Silicon Carbide

- [Materials for Mechanical Seals](#)

Semi-analytical EHL Solution Methods

- [Simplified EHL Solution Methods](#)

Semi-infinite Body Deformation Analyzed with the Differential Deflection Method

- [Differential Deflection Method for Deformation](#)

Semi-solid Lubricant

- [Lubricating Grease](#)

Semi-system Approach

WEN-ZHONG WANG

School of Mechanical Engineering, Beijing Institute of Technology, Beijing, People's Republic of China

Definition

The elasticity equation and the Reynolds equation are solved separately and sequentially, and pressures are updated from both the pressure flow term and the shear flow term of Reynolds equation.

Scientific Fundamentals

Elastohydrodynamic lubrication (EHL) is a common lubrication condition in gears, bearings, cams, and traction drivers. The EHL problem can be well formulated by Reynolds equation, which describes the formation of hydrodynamic pressure in lubricated contact, the film thickness equation, and the load balance equation, plus two equations of state relating the physical properties of the lubricant to the pressures generated inside the contact for isothermal conditions. These equations are common and can be found in literature related to lubrication analysis. Their common forms for point-contact are

$$\begin{aligned} \text{Reynolds equation: } & \frac{\partial}{\partial x} \left(\frac{\rho h^3}{12\eta} \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial y} \left(\frac{\rho h^3}{12\eta} \frac{\partial p}{\partial y} \right) \\ & = V_r \frac{\partial(\rho h)}{\partial x} + \frac{\partial(\rho h)}{\partial t} \end{aligned} \quad (1)$$

$$\begin{aligned} \text{Film thickness equation: } & h(x, y, t) = h_0 + \frac{x^2}{2R_x} + \frac{y^2}{2R_y} \\ & + v(x, y, t) + \delta(x, y, t) \end{aligned} \quad (2)$$

$$\text{Load balance equation: } \int_{x_0}^{x_e} \int_{y_0}^{y_e} p(x, y, t) dx dy = W \quad (3)$$

$$\text{Surface elastic deformation: } v(x, y) = \frac{2}{\pi E^*} \times \iint_{\Omega} \frac{p(\xi, \zeta)}{\sqrt{(x-\xi)^2 + (y-\zeta)^2}} d\xi d\zeta \quad (4)$$

Introduce the following nondimensional parameter groups:

$$P = \frac{p}{p_H}, H = \frac{h}{a}, X = \frac{x}{a}, Y = \frac{y}{b}, \bar{\eta} = \frac{\eta}{\eta_0}, \bar{\rho} = \frac{\rho}{\rho_0}, \bar{t} = \frac{tV_r}{a}$$

where p_H , a , b are the Hertzian dry contact parameters, and η_0 and ρ_0 are the ambient viscosity and density of lubricant, respectively.

Based on the nondimensional parameter groups, the Reynolds equation and film thickness equation can be normalized in the following nondimensional forms for the steady state condition, for example,

$$\frac{\partial}{\partial X} \left(\varepsilon^x \frac{\partial P}{\partial X} \right) + \frac{\partial}{\partial Y} \left(\varepsilon^y \frac{\partial P}{\partial Y} \right) = \frac{\partial(\bar{\rho}H)}{\partial X} \quad (5)$$

$$H = H_0 + B_x X^2 + B_y Y^2 + \bar{\delta}(X, Y) + V(X, Y) \quad (6)$$

$$\text{where } \varepsilon^x = \left(\frac{ap_h}{12V_r\eta_0} \right) \left(\frac{\bar{\rho}H^3}{\bar{\eta}} \right), \quad \varepsilon^y = \frac{\varepsilon^x}{K_e^2}, \quad B_x = \frac{a}{2R_x}, \\ B_y = \frac{aK_e^2}{2R_y}$$

The solution of this set of equations provides predictions of film thickness and pressures in lubricated contacts. In the development of EHL theory, the full numerical method plays a very significant role. In the early development of EHL, the Gauss-Seidal iteration and Newton-Raphson method were used to solve light-load lubrication problems. With progress in science and technology, an increasing number of mechanical components will operate under severe conditions. The lubrication problems encountered are more complicated than ever before. Most of the numerical methods and algorithms suffer the convergent difficulty under the severe conditions. The main reason for the numerical instability encountered in the above-mentioned methods must be understood in order to develop more robust numerical algorithms. Under heavily loaded conditions, lubricant in a contact zone is pressured and the viscosity increases almost exponentially with the pressure. Near the contact center, the viscosity of the lubricant can be several orders higher than the ambient value. As a result, the Poiseuille term standing for pressure flow as expressed by the two left

terms in the Reynolds equation cease to play the dominant role and, instead, the shear flow term and squeezing term as expressed by the two right terms of the Reynolds equation take over. The change of roles of pressure flow and shear flow as the applied load increases implies the possibility of development of a robust algorithm for EHL numerical simulations.

By using the appropriate differential schemes, the Reynolds equation (5) can be converted into a discrete differential equation at each unknown pressure point. A typical form for a discrete equation is a three-point stencil as follows:

$$A_{i,j}P_{i-1,j} + B_{i,j}P_{i,j} + C_{i,j}P_{i+1,j} = F_{i,j} \quad (7)$$

For pressure flow terms, the second central differential scheme is commonly used. By substitution, the expression for pressure flow terms becomes

$$\left[\frac{\partial}{\partial X} \left(\varepsilon^x \frac{\partial P}{\partial X} \right) \right]_{i,j} = \frac{1}{\Delta X^2} [\varepsilon_{i+1/2,j}^x P_{i+1,j} - (\varepsilon_{i+1/2,j}^x + \varepsilon_{i-1/2,j}^x) P_{i,j} + \varepsilon_{i-1/2,j}^x P_{i-1,j}] \quad (8)$$

$$\left[\frac{\partial}{\partial Y} \left(\varepsilon^y \frac{\partial P}{\partial Y} \right) \right]_{i,j} = \frac{1}{\Delta Y^2} [\varepsilon_{i,j+1/2}^y P_{i,j+1} - (\varepsilon_{i,j+1/2}^y + \varepsilon_{i,j-1/2}^y) P_{i,j} + \varepsilon_{i,j-1/2}^y P_{i,j-1}] \quad (9)$$

where

$$\varepsilon_{i+1/2,j}^x = \frac{1}{2} (\varepsilon_{i,j}^x + \varepsilon_{i+1,j}^x), \quad \varepsilon_{i-1/2,j}^x = \frac{1}{2} (\varepsilon_{i-1,j}^x + \varepsilon_{i,j}^x) \\ \varepsilon_{i,j+1/2}^y = \frac{1}{2} (\varepsilon_{i,j}^y + \varepsilon_{i,j+1}^y), \quad \varepsilon_{i,j-1/2}^y = \frac{1}{2} (\varepsilon_{i,j-1}^y + \varepsilon_{i,j}^y)$$

There are several differential schemes that can be used to discretize the shear flow term. For example, if the first backward scheme is used for the shear flow term (wedge term), the expression for shear flow term becomes

$$\left[\frac{\partial(\bar{\rho}H)}{\partial X} \right]_{i,j} = \frac{\bar{\rho}_{i,j}H_{i,j} - \bar{\rho}_{i-1,j}H_{i-1,j}}{\Delta X} \quad (10)$$

Substituting (8), (9), and (10) into (5), one has,

$$\frac{\varepsilon_{i+1/2,j}^x P_{i+1,j} - (\varepsilon_{i+1/2,j}^x + \varepsilon_{i-1/2,j}^x) P_{i,j} + \varepsilon_{i-1/2,j}^x P_{i-1,j}}{\Delta X^2} + \frac{\varepsilon_{i,j+1/2}^y P_{i,j+1} - (\varepsilon_{i,j+1/2}^y + \varepsilon_{i,j-1/2}^y) P_{i,j} + \varepsilon_{i,j-1/2}^y P_{i,j-1}}{\Delta Y^2} \\ = \frac{\bar{\rho}_{i,j}H_{i,j} - \bar{\rho}_{i-1,j}H_{i-1,j}}{\Delta X} \quad (11)$$

Rearrange (11) into the form of (7), where the coefficients are as follows, respectively,

$$\begin{cases} A_{ij} = \frac{\varepsilon_{i-1/2,j}^x}{\Delta X^2} \\ B_{ij} = -\frac{\varepsilon_{i-1/2,j}^x + \varepsilon_{i+1/2,j}^x}{\Delta X^2} - \frac{\varepsilon_{i,j-1/2}^y + \varepsilon_{i,j+1/2}^y}{\Delta Y^2} \\ C_{ij} = \frac{\varepsilon_{i+1/2,j}^x}{\Delta X^2} \\ F_{ij} = \frac{\bar{\rho}_{ij}H_{ij} - \bar{\rho}_{i-1,j}H_{i-1,j}}{\Delta X} - \frac{\varepsilon_{i,j-1/2}^y \bar{P}_{i,j-1}}{\Delta Y^2} - \frac{\varepsilon_{i,j+1/2}^y \bar{P}_{i,j+1}}{\Delta Y^2} \end{cases} \quad (12)$$

In the Gauss-Seidal iteration, F_{ij} is treated as known variables, which means the variables H , ρ , ε , appearing in the expression of F_{ij} are evaluated by the old pressure approximations. For given approximation P_{ij} , the nodal points are visited in the lexicographic order. At each grid point a new approximation \tilde{P}_{ij} is updated according to

$$A_{ij}\tilde{P}_{i-1,j} + B_{ij}\tilde{P}_{i,j} + C_{ij}\tilde{P}_{i+1,j} = F_{ij} \quad (13)$$

where pressure with “~” is the new estimate of solution, and F_{ij} is treated as known variable calculated from the old approximation.

The iterative equation (13) often encounters convergence difficulty under heavily loaded conditions. As mentioned, physically, the shear flow term plays a very important role under heavily loaded conditions, thus, when constructing iterative formula, incorporating the contribution from the shear flow term may help to regain the numerical stability. Mathematically, with the applied load increasing gradually, the coefficient matrix used in the Gauss-Seidal iteration process constructed only from the pressure flow term will gradually lose the characteristic that the leading diagonal element is dominant, eventually becoming an ill-conditioned matrix, which results in poor convergence property. If the shear flow term (dH/dx) is expressed by the unknown pressure at each nodal point, and its contribution is considered when constructing the coefficient matrix, the convergence process will be improved. Thus, in each iteration loop the pressures appearing in the left and right of the Reynolds equation are updated simultaneously. This is called the *semi-system approach* (Ai 1993). The following example shows the implementation of the semi-system method to the steady-state point-contact lubrication problem (Liu et al. 2005).

The second central differential scheme is still applied to the pressure flow term; the coefficient contributions of

the pressure flow term to discrete equation (7) can be obtained using the following expressions:

$$\begin{cases} A_{ij}^p = \frac{\varepsilon_{i-1/2,j}^x}{\Delta X^2} \\ B_{ij}^p = -\frac{\varepsilon_{i+1/2,j}^x + \varepsilon_{i-1/2,j}^x}{\Delta X^2} - \frac{\varepsilon_{i,j+1/2}^x + \varepsilon_{i,j-1/2}^x}{\Delta Y^2} \\ C_{ij}^p = \frac{\varepsilon_{i+1/2,j}^x}{\Delta X^2} \\ F_{ij}^p = -\frac{\varepsilon_{i,j+1/2}^y P_{i,j+1} + \varepsilon_{i,j-1/2}^y P_{i,j-1}}{\Delta Y^2} \end{cases} \quad (14)$$

Coefficient B corresponds to the leading diagonal element of coefficient matrix. It is negative and has the largest absolute value compared with other coefficients A and C .

The first backward scheme as an example is applied to the shear term, and the expression is the same as (10). Because film thicknesses at two nodal points, (i, j) and $(i-1, j)$, are related to all pressures, in a discrete form, these film thicknesses can be expressed as

$$H_{ij} = H_0 + B_x X_{ij}^2 + B_y Y_{ij}^2 + \sum_k \sum_l D_{k,l}^{ij} P_{k,l} \quad (15)$$

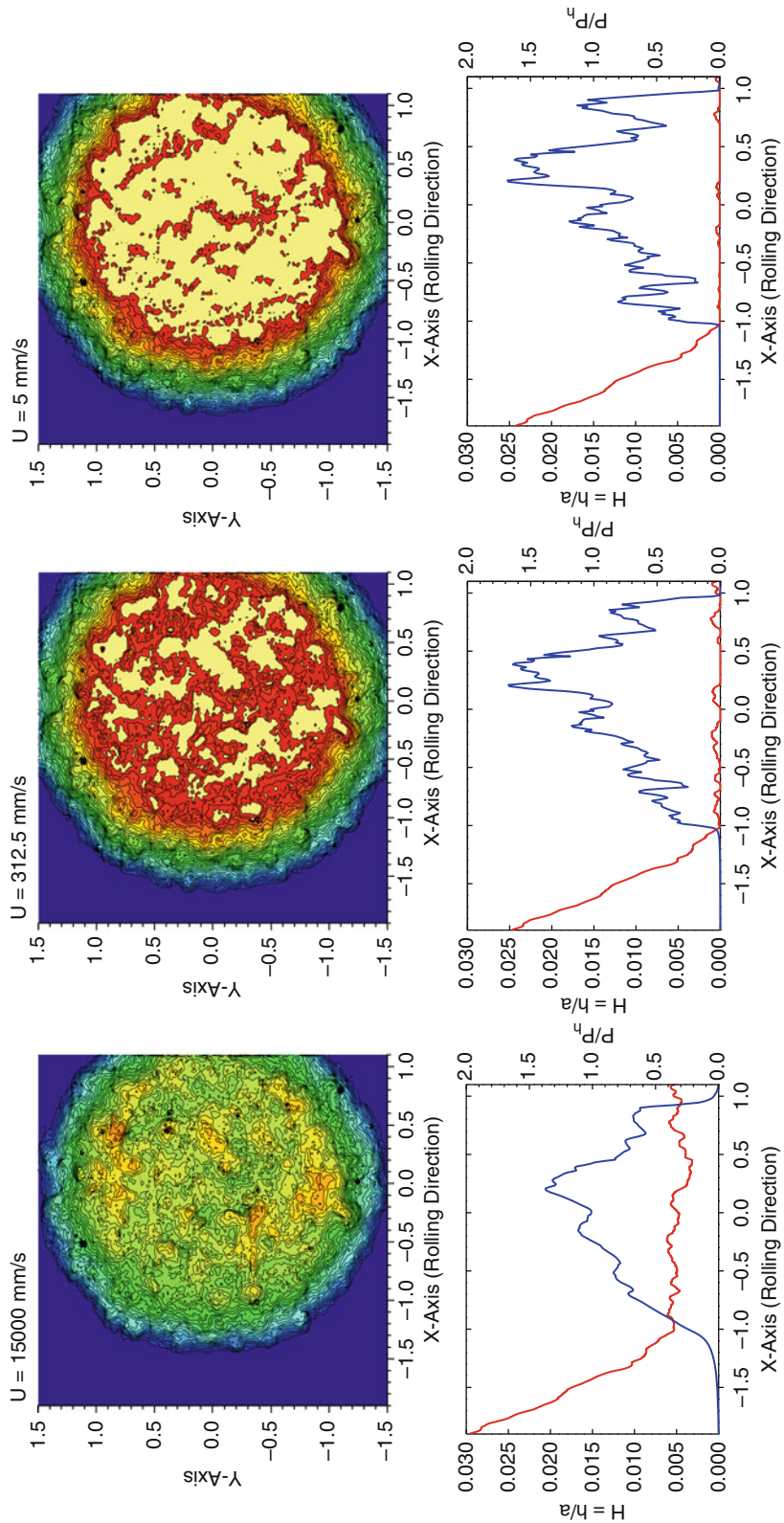
$$H_{i-1,j} = H_0 + B_x X_{i-1,j}^2 + B_y Y_{i-1,j}^2 + \sum_k \sum_l D_{k,l}^{i-1,j} P_{k,l} \quad (16)$$

where $D_{k,l}^{ij}$ is the influence coefficient relating to the normal deformation at point (x_i, y_j) owing to a unit load acting on point (x_k, y_l) . It is obviously that $D_{k,l}^{k,l}$, the deformation at the point where the pressure is applied, is the largest.

By substituting (15) and (16) into (10), the coefficient contributions from shear flow term to discrete equation (7) can be obtained as follows:

$$\begin{cases} A_{ij}^w = -\frac{\bar{\rho}_{ij}D_{i-1,j}^{ij} - \bar{\rho}_{i-1,j}D_{i-1,j}^{i-1,j}}{\Delta X} \\ B_{ij}^w = -\frac{\bar{\rho}_{ij}D_{i,j}^{ij} - \bar{\rho}_{i-1,j}D_{i,j}^{i-1,j}}{\Delta X} \\ C_{ij}^w = -\frac{\bar{\rho}_{ij}D_{i+1,j}^{ij} - \bar{\rho}_{i-1,j}D_{i+1,j}^{i-1,j}}{\Delta X} \\ F_{ij}^w = \frac{1}{\Delta X} \left[\bar{\rho}_{ij}[H_{ij} - (D_{i-1,j}^{ij} P_{i-1,j}^{old} + D_{i,j}^{ij} P_{i,j}^{old} + D_{i+1,j}^{ij} P_{i+1,j}^{old})] \right. \\ \left. - \bar{\rho}_{i-1,j}[H_{i-1,j} - (D_{i-1,j}^{i-1,j} P_{i-1,j}^{old} + D_{i,j}^{i-1,j} P_{i,j}^{old} + D_{i+1,j}^{i-1,j} P_{i+1,j}^{old})] \right] \end{cases} \quad (17)$$

As the density variation of adjacent points is relatively smaller, but the influence coefficient $D_{i,j}^{i,j}$ at the current point is the largest, it is evident that the coefficient B in (17) is negative and has the largest absolute value.



Semi-system Approach, Fig. 1 The EHL numerical solutions at different velocities with measured surface

The total coefficients of discrete equation (7) are summations of the coefficients from pressure flow term and shear flow term. That is,

$$\begin{cases} A_{i,j} = A_{i,j}^P + A_{i,j}^W \\ B_{i,j} = B_{i,j}^P + B_{i,j}^W \\ C_{i,j} = C_{i,j}^P + C_{i,j}^W \\ F_{i,j} = F_{i,j}^P + F_{i,j}^W \end{cases} \quad (18)$$

After assembling the contributions of the two sides, the leading diagonal element of coefficient matrix has the largest absolute value. This ensures the convergence of relaxation iteration based on the dominance theorem. The computation experience shows that an under-relaxation factor of 0.02 is appropriate for this scheme. In fact, when the semi-system approach is applied to the EHL problem, there are other differential schemes for the shear flow term appearing in the Reynolds equation, by choosing appropriate differential schemes, the convergence can be further improved.

Key Application

As the semi-system approach can ensure the leading diagonal element of the coefficient matrix is dominant, the convergence of relaxation iteration when solving the Reynolds equation is good enough. Thus, the semi-system approach can be extensively used in the analysis of EHL problem, especially for heavy-duty conditions. One of the successful applications of the semi-system approach is the unified Reynolds equation model for mixed lubrication proposed by (Hu and Zhu 2000). Based on this approach, this model can handle the entire transition from full-film and mixed EHL down to boundary lubrication by solving hydrodynamic and contact pressures simultaneously, which means one can tackle lubrication problems in an extended film thickness range (i.e., from several microns down to a few nanometers and even to practical zero). Actually, as the entraining speed continuously decreases, the solutions obtained from the EHL equation system show that the hydrodynamic effect gradually vanishes and the pressure distribution and deformed contact geometry approach those of dry contact. In all the cases, the core iteration algorithm, the semi-system approach, shows very good convergence property and does not encounter any convergence problems, even if severe contacts occur. Figure 1 shows the EHL numerical solutions at different velocities with measured surface, which is based on the semi-system approach and implies the performance of the approach.

Nomenclature

a, b	Hertz contact radius in the x and y directions (m)
$D_{k,l}^{i,j}$	Deformation influence coefficient at point (i, j) owing to a unit load acting on positing (k, l)
$A_{i,j}, B_{i,j}, C_{i,j}, F_{i,j}$	Coefficients for discrete Reynolds equation
E^*	Effective Young's modulus (Pa)
h, H	Film thickness and nondimensional film thickness
h_0, H_0	Normal approach of two stiff surface (m) and dimensionless normal approach, $H_0 = h_0/a$
K_e	Ellipticity, $K_e = b/a$
R_x, R_y	Reduced radius of curvature in x and y direction
V_r	$= (V_1 + V_2)/2$, Entrainment velocity (m/s)
η, η_0	Viscosity and ambient viscosity of lubricant (Pa. s)
ρ, ρ_0	Density and ambient density of lubricant (kg/m^3)
i, j	Position (x_i, y_j)
W	Applied load (N)
t, \bar{t}	Dimensional (s) and nondimensional time
p	Contact pressure (Pa)
p_H	Maximum Hertzian pressure (Pa)
v	Normal surface deformation
x, y, X, Y	Coordinate in the x and y direction and nondimensional counterparts
x_0, x_e	Inlet and outlet distance in x direction
y_0, y_e	Inlet and outlet distance in y direction
$\Delta x, \Delta y$	Dimensions of an element in the x and y direction
δ	Surface roughness height
Tidal($^\circ$)	New approximation of variables

Cross-References

- 3D Line Contact EHL
- Differential Scheme Effect on EHL Solution
- EHL Governing Equations
- EHL, Full Numerical Solution Methods
- EHL-Line Contact
- Lubricant Non-Newtonian Effect on EHL
- Mesh Density Effect on EHL Solution
- Mixed EHL
- Plasto-Elastohydrodynamic Lubrication (PEHL)
- Point Contact EHL

- [Surface Deformation Calculation for EHL](#)
- [Thermal EHL Theory](#)

References

- X.L. Ai, Numerical analysis of elastohydrodynamically lubricated line and point contacts with rough surfaces by using semi-system and multigrid methods. Ph.D. Thesis, Northwestern University, IL, 1993
- Y.Z. Hu, D. Zhu, A full numerical solution to the mixed lubrication in point contacts. *J. Tribol. Trans. ASME* **122**, 1–9 (2000)
- Y.C. Liu, Q. Wang, W.Z. Wang, Y.Z. Hu, D. Zhu, Effects of differential scheme and mesh density on EHL film thickness in point contacts. *J. Tribol. Trans. ASME* **128**, 641–653 (2005)

Sesame Oil

- [Natural Oils as Lubricants](#)

Severe Adhesion

- [Gear Contact Temperature and Scuffing Risk Analysis](#)

SFA-Surface Force Apparatus

- [Surface Force Apparatus](#)

SFE–Surface Free Energy

- [Surface Free Energy](#)

SFFT – Spherical Fast Fourier Transform Considering Spherical/Aspheric Geometry

- [Geometry of Spherical/Aspheric Bearings](#)

SFFT – Spherical Fast Fourier Transform For Joint Simulation

- [Biotribological Joint Simulation System](#)

SFFT – Spherical Fast Fourier Transform for Spherical-Bearing Friction Prediction

- [Friction Prediction for Spherical Bearings](#)

SFFT – Spherical Fast Fourier Transform for Spherical-Bearing Lubrication Analysis

- [Lubrication Theory for Spherical Bearings](#)

SFFT – Spherical Fast Fourier Transform for Spherical-Bearing Wear Prediction

- [Wear Modeling of Spherical Bearings](#)

S_f-N Relationships for Ferrous Materials

- [Fatigue Strength-Load Cycle Relationships for Ferrous Materials](#)

SGDM – Spherical Grid Data Model Considering Spherical/Aspheric Geometry

- [Geometry of Spherical/Aspheric Bearings](#)

SGDM – Spherical Grid Data Model for Joint Simulation

- [Biotribological Joint Simulation System](#)

SGDM – Spherical Grid Data Model for Spherical-Bearing Friction Prediction

- [Friction Prediction for Spherical Bearings](#)

SGDM – Spherical Grid Data Model for Spherical-Bearing Lubrication Analysis

- [Lubrication Theory for Spherical Bearings](#)

SGDM – Spherical Grid Data Model for Spherical-Bearing Wear Modeling

- [Wear Modeling of Spherical Bearings](#)

Shaft Packing

- [Compression Packing to Form Seals for Rotating Shafts](#)

Shaft Seal

- [Rotary Shaft Lip Seals](#)

Shaft Seal Materials

- [Lip Seals, Materials](#)

Shaft Seal Methods

- [Mechanical Seals](#)

Shakedown

PHAM DUC CHINH
VAST, Hanoi, Vietnam

Definition

A structure made of elastic plastic material in a given loading scheme, after an initial stage of possible limited plastic deformation (of finite total plastic dissipation), may eventually *shake down* to some residual stress state, from which it subsequently responds elastically (and, hence, safely) to the external agencies. Otherwise, the structure is considered as having failed, because the plastic deformation would accumulate unrestrictively (the mode is called ratcheting or incremental collapse), or be bounded but vary cyclically and unceasingly (fatigue, cyclic, or alternating plasticity collapse mode).

Scientific Fundamentals

In principle, shakedown incremental analysis can apply to any sophisticated elastic plastic material model in small or large deformation, following a particular loading history. Powerful shakedown theorems of Melan-Koiter type (Koiter 1963; König 1987; Pham 2003a) can be constructed for generally kinematic hardening bodies (Pham 2007, 2008) within the framework of classical plasticity theory with small deformation and normality flow rule assumptions. The essential advantage of the theorems is their path-independence: the theorems determine the time-independent boundary in the loading space, under which a structure is safe regardless of particular loading histories, while the structure fails if the boundary is violated unrestrictively. Like their limiting case – the plastic limit theorems, and the minimum energy and complementary energy principles of elasticity – the shakedown static and kinematic theorems are stated as optimization problems. With any trial admissible static field the static theorem gives a lower bound estimate for the shakedown load limit, while with a trial admissible kinematic field the kinematic theorem yields an upper bound one (the optimal static as well as kinematic fields determine the exact shakedown limit).

Let $\sigma^e(\mathbf{x}, t)$ denote the fictitious elastic stress response of the body V (under the assumption of its perfectly elastic behavior) to external agencies over a period of time ($\mathbf{x} \in V, t \in [0, T]$), called a loading history. The actions of all kinds of external agencies upon V can be expressed explicitly through σ^e . At every point $\mathbf{x} \in V$, the elastic stress response $\sigma^e(\mathbf{x}, t)$ is confined to a bounded

time-independent domain with prescribed limits in the stress space, called a local loading domain \mathcal{L}_x . As a field over V , $\boldsymbol{\sigma}^e(\mathbf{x}, t)$ belongs to the time-independent global loading domain \mathcal{L} :

$$\mathcal{L} = \{\boldsymbol{\sigma}^e | \boldsymbol{\sigma}^e(\mathbf{x}, t) \in \mathcal{L}_x, \mathbf{x} \in V, t \in [0, T]\}. \quad (1)$$

In the spirit of shakedown theorems, the bounded loading domain \mathcal{L} , instead of a particular loading history $\boldsymbol{\sigma}^e(\mathbf{x}, t)$, is given a priori. Shakedown of a body in \mathcal{L} means it shakes down for all possible loading histories $\boldsymbol{\sigma}^e(\mathbf{x}, t) \in \mathcal{L}$.

Originally Melan-Koiter shakedown theorems have been constructed for idealistic elastic perfectly plastic material model, which is practical and suffice for plastic limit analysis. For shakedown analysis, much more realistic and practical material model is the elastic plastic kinematic hardening one involving Bauschinger effect. The plastic hardening curve is generally nonlinear, plastic-deformation-history dependent, and is bounded by the initial and ultimate yield stresses. Shakedown static and kinematic theorems have been constructed for the kinematic hardening material satisfying a realistic positive hysteresis postulate, which do not involve the plastic-deformation-history dependent hardening curve, except the initial yield stress and ultimate yield strength, keeping the path-independent spirit of classical Melan-Koiter theorems (Pham 2007, 2008).

Let k_s denote the shakedown safety factor: at $k_s > 1$ the structure will shake down, while it will not at $k_s < 1$, and $k_s = 1$ defines the boundary of the shakedown domain.

Shakedown Static Theorem

$$k_s = \min \{\bar{U}, \bar{C}\}, \quad (2)$$

where

$$\bar{U} = \sup_{\boldsymbol{\rho} \in \mathcal{R}} \{k | k(\boldsymbol{\rho} + \boldsymbol{\sigma}^e) \in \mathcal{Y}_u, \forall \boldsymbol{\sigma}^e \in \mathcal{L}\}, \quad (3)$$

$$\bar{C} = \sup_{\boldsymbol{\rho}'} \{k | k(\boldsymbol{\rho}' + \boldsymbol{\sigma}^e) \in \mathcal{Y}_i, \forall \boldsymbol{\sigma}^e \in \mathcal{L}\}, \quad (4)$$

\mathcal{R} is the set of admissible time-independent, self-equilibrated residual stress fields $\boldsymbol{\rho}(\mathbf{x})$ that satisfy homogeneous equilibrium equations on V ; $\boldsymbol{\rho}'$ is a time-independent stress field that is not required to be self-equilibrated; \mathcal{Y}_u designates the elastic domain in the stress space that is bounded by the yield surface determined by the ultimate yield stress σ_Y^u , while \mathcal{Y}_i is the respective domain bounded by the yield surface determined by the initial yield stress σ_Y^i .

In the case $\sigma_Y^i = \sigma_Y^u = \sigma_Y$, statement (3) leads to the classical shakedown static theorem for perfectly plastic body: at $\bar{U} > 1$ (safe), sum of the time-independent residual stress $\boldsymbol{\rho}$ and the elastic stress $\boldsymbol{\sigma}^e$ over the whole body is in a safe state defined by the yield surface \mathcal{Y}_u ; while at $\bar{U} < 1$ (unsafe) no such residual stress should exist. Statement (4) verifies the possibility of the bounded cyclic plasticity mode determined by the yield stress σ_Y^i : if the range of the varying part of the stress $\boldsymbol{\sigma}^e$ (from the static part $\boldsymbol{\rho}'$) everywhere inside the body should be smaller than the size of the yield surface \mathcal{Y}_i (safe with respect to the mode), or not (unsafe).

Shakedown Kinematic Theorem

$$k_s^{-1} = \max \{U, C\}, \quad (5)$$

where

$$U = \sup_{\mathbf{e}^p \in \mathcal{A}; \boldsymbol{\sigma}^e \in \mathcal{L}} \frac{\int_0^T dt \int_V \boldsymbol{\sigma}^e : \mathbf{e}^p dV}{\int_0^T dt \int_V D_u(\mathbf{e}^p) dV}, \quad (6)$$

$$C = \sup_{\mathbf{x} \in V; \boldsymbol{\sigma}^e \in \mathcal{L}; \hat{\mathbf{e}}^p : \boldsymbol{\rho}'} \frac{(\boldsymbol{\sigma}^e + \boldsymbol{\rho}') : \hat{\mathbf{e}}^p}{D_i(\hat{\mathbf{e}}^p)}, \quad (7)$$

\mathcal{A} is the set of compatible-end-cycle plastic strain rate fields \mathbf{e}^p over the time cycles $0 \leq t \leq T$:

$$\mathcal{A} = \left\{ \mathbf{e}^p \mid \boldsymbol{\varepsilon}^p = \int_0^T \mathbf{e}^p dt \in \mathcal{C} \right\}; \quad (8)$$

\mathcal{C} is the set of compatible plastic strain increment fields on V ; $\hat{\mathbf{e}}^p$ and $\boldsymbol{\rho}'$ are plastic strain rate- and time-independent stress fields that are not required to satisfy any compatibility and equilibrium constraints; $D(\mathbf{e}^p)$ is the dissipation function determined by the yield stress σ_Y and the respective yield criterion, for example, for a von Mises material:

$$D(\mathbf{e}^p) = \sqrt{2/3} \sigma_Y (\mathbf{e}^p : \mathbf{e}^p)^{1/2}, \quad (9)$$

while for a Tresca material:

$$D(\mathbf{e}^p) = \frac{1}{2} \sigma_Y (|e_1^p| + |e_2^p| + |e_3^p|) \\ = \sigma_Y \max\{|e_1^p|, |e_2^p|, |e_3^p|\}, \quad (10)$$

e_1^p, e_2^p, e_3^p are the principal plastic strain rates; $D_u(\mathbf{e}^p)$ and $D_i(\mathbf{e}^p)$ are the dissipation functions with σ_Y^u and σ_Y^i taking the places of σ_Y respectively.

In the case $\sigma_Y^i = \sigma_Y^u = \sigma_Y$, statement (6) leads to the classical shakedown kinematic theorem for perfectly plastic body: at $U < 1$ (safe), the internal plastic dissipation

capacity of the body greater than the possible mechanical work of the external agencies; while at $U > 1$ (unsafe) the reverse is true. Statement (7) verifies the possibility of the bounded cyclic plasticity mode determined by the yield stress σ_Y^i : if the range of the varying part of the stress σ^e (from the static part ρ') everywhere inside the body should be smaller than the size of the yield surface \mathcal{Y}_i (safe with respect to the mode), or not (unsafe).

In summary, statements (3) and (6) are identical to those of the shakedown static and kinematic theorems for elastic perfectly plastic material with the (ultimate) yield stress σ_Y^u , while statements (4) or (7) are just expressions of the bounded cyclic plasticity mode determined by the (initial) yield stress σ_Y^i . Note that, in contrast to the global mode (3) involving the self-equilibrated residual stress field ρ over V [or (6) involving the end-cycle-compatible plastic strain rate field ϵ^p over V], the mode (4) [or (7)] is local and can be checked at every point $\mathbf{x} \in V$ separately. At $\bar{U} > \bar{C}$ of criterion (2) [or $U < C$ of criterion (5)] the non-shakedown collapse mode is cyclic plasticity, otherwise the incremental collapse mode prevails.

For applications, the following reduced kinematic theorem is useful (Pham and Stumpf 1994; Pham 2008):

Reduced Kinematic Theorem

$$k_s^{-1} \geq \hat{k}_s^{-1} = \max\{I, A\}, \quad (11)$$

where

$$I = \sup_{\sigma^e \in \mathcal{L}; \epsilon^p \in \mathcal{C}} \frac{\int_V \max_{t_x} [\sigma^e(\mathbf{x}, t_x) : \epsilon^p(\mathbf{x})] dV}{\int_V D_u(\epsilon^p) dV} \quad (12)$$

$$A = \sup_{\mathbf{x} \in V; \sigma^e \in \mathcal{L}; \hat{\epsilon}^p; t_1, t_2} \frac{[\sigma^e(\mathbf{x}, t_1) - \sigma^e(\mathbf{x}, t_2)] : \hat{\epsilon}^p(\mathbf{x})}{2D_i(\hat{\epsilon}^p)}. \quad (13)$$

The reduced theorem (11)–(13) is simpler than the theorem (5)–(7), in particular as the incremental collapse criterion (12) is compared to the criterion (6): the incremental collapse mode (12) does not involve time integrals and has the expression almost as simple as that of the respective plastic limit kinematic theorem, with only the difference that the underintegral maximum operation over time parameter t is taken at every point $\mathbf{x} \in V$ separately, which makes the incremental collapse criterion (12) more conservative than the plastic limit one. Hence, available kinematic methods of plastic limit analysis can be modified to be used to solve problem (12). Meanwhile, the simple (13) expresses the alternating plasticity collapse

mode. For a broad class of practical problems, where the components of plastic deformations at every point inside a structure should change proportionally during loading cycles, the exact equality $k_s = \hat{k}_s$ has been proved. Generally, \hat{k}_s is expected to provide a good upper bound estimate, if not the exact value, of k_s (no particular example has been found such that the strict inequality $k_s < \hat{k}_s$ holds).

Key Applications

In the following, we consider some examples of application of (11)–(13), which yield analytical and semi-analytical solutions. In the examples, we have $A = C = \bar{C}$; and $I = U$ ($I = \bar{U}$) whenever $U > C = A$ ($\bar{U} < \bar{C} = A$). Simple examples of application of the shakedown static theorem can be consulted, for example, in (König 1987). Generally, the stated nonlinear mathematical programming problems (2)–(4), (5)–(7), or (11)–(13) should be solved by numerical methods. See more about applications in (Gokhfeld and Cherniavski 1980; König 1987; Pham and Stumpf 1994; Pham 2000, 2003b, 2008, 2010; Weichert and Maier 2002; Tran et al. 2010; and the references therein)

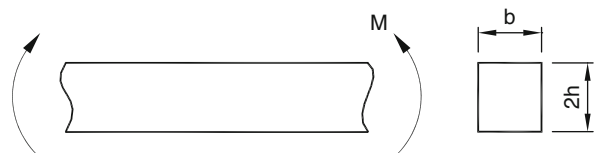
Examples

Example 1. A beam of rectangular cross-section of the size $b \times 2h$ is bent up and down alternatively with the moment $\pm M$ (Fig. 1). The outermost layers on the upper and lower sides of the beam begin to yield at the moment $M = M_Y^i = \sigma_Y^i \frac{2}{3}bh^2$, while the ultimate yielding moment (the plastic limit load) of the beam is $M = M_Y^u = \sigma_Y^u bh^2$. Application of (11)–(13) gives the obvious result:

$$k_s^{-1} = \max\{I, A\} = \max\{M/M_Y^u, M/M_Y^i\} = M/M_Y^i; \quad (14)$$

hence at $M = M_Y^i$ ($k_s = 1$), the beam fails because of alternating plasticity started from the failure of the beam's outermost layers.

Example 2. Consider a beam clamped at the two ends A and D (no horizontal kinematic constraint). The vertical

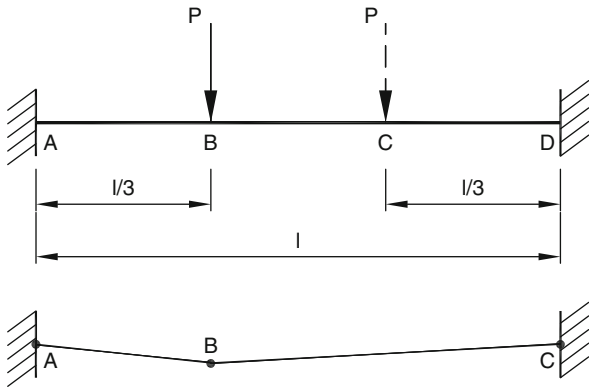


Shakedown, Fig. 1 A beam in bending

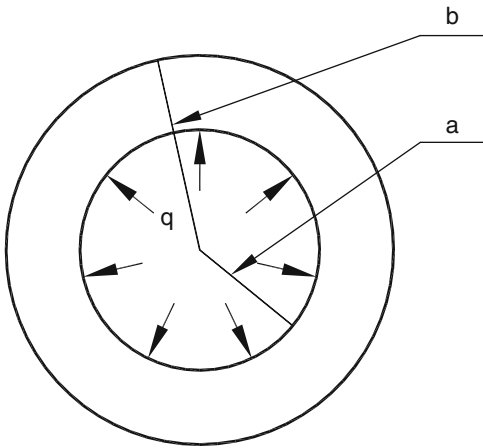
point load P is applied and removed slowly, but alternatively, at B and then C for an unlimited number of times. Application of criterion (12) with an optimal incremental collapse mechanism ABC shown in Fig. 2 yields the non-shakedown limit load: $P_{SD} = 8.1M_Y^u/l$ (the respective plastic limit load is significantly larger: $P_p = 9M_Y^u/l$).

Example 3. A thick-walled hollow sphere of inner and outer radii a and b is subjected to quasistatic internal pressure q , which may vary arbitrarily from 0 to the limit q^U (Fig. 3). Application of criterion (11)–(13) yields

$$k_s^{-1} = \max \{I, A\} = q^U \max \left\{ \frac{1}{2\sigma_Y^u \ln(b/a)}, \frac{3b^3}{4\sigma_Y^i (b^3 - a^3)} \right\}, \quad (15)$$



Shakedown, Fig. 2 A clamped beam under sequential loads



Shakedown, Fig. 3 A hollow sphere under variable pressure

that is, the non-shakedown pressure limit (at $k_s = 1$):

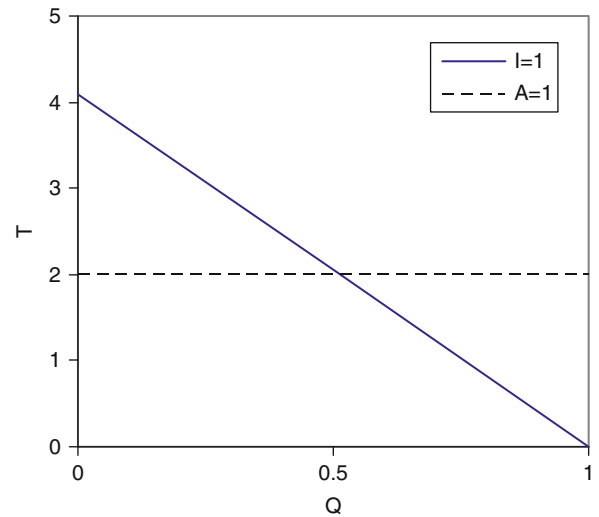
$$q^U = q_{SD} = \min \left\{ \sigma_Y^u 2 \ln(b/a), \sigma_Y^i \frac{4(b^3 - a^3)}{3b^3} \right\}. \quad (16)$$

Example 4. A thin cylinder shell of inner and outer radii a and b is subjected to constant internal pressure q and quasistatic internal temperature θ , which may vary arbitrarily from 0 to the shakedown limit (the reference external ambient temperature is taken as zero). The cylinder is long and its ends are closed with rigid diaphragms. The incremental collapse line $I = 1$ and the alternating plasticity one $A = 1$ are presented in the plane of dimensionless external load ($Q = \frac{q}{\sigma_Y^i \ln(b/a)}$) and temperature ($T = \theta \frac{E\alpha}{2(1-\nu)\sigma_Y^i}$) parameters in Fig. 4, where E and ν are the Young's modulus and Poisson's ratio, α – thermal expansion coefficient. For the numerical illustration, $\frac{b}{b-a} = 20$ and $\sigma_Y^u = \sigma_Y^i$ (perfect plasticity) are given. The domain enveloped above by both lines is the shakedown domain.

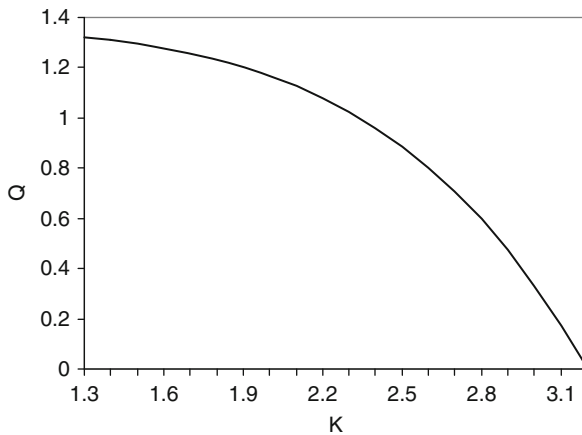
Example 5. A disk of radius a , thickness h , and mass density m , clamped at the edge, is subjected from one side to uniform quasiperiodic dynamic pressure

$$q = q_0 + q_1 \sin \omega t, \quad (17)$$

where q_0 is constant, while q_1 and ω are quasistatic functions of time varying between the limits



Shakedown, Fig. 4 Shakedown domain and collapse modes for the cylinder shell. Incremental ($I = 1$) and alternating plasticity ($A = 1$) lines in the plane of external load and temperature parameters



Shakedown, Fig. 5 Non-shakedown curve for the disk under dynamic loading

$$0 \leq q_1 \leq q_1^U, \quad 0 \leq \omega \leq \omega_U < \omega_I, \quad (18)$$

ω_I is the principal natural frequency of the structure. Denote $k^A = \frac{hm}{D} \omega^2$, $k_U^A = \frac{hm}{D} \omega_U^2$, $D = \frac{Elr^3}{12(1-\nu^2)}$, $M_Y^u = \frac{h^2}{4} \sigma_Y^u$, $M_Y^i = \frac{h^2}{4} \sigma_Y^i$. For numerical illustration, take $q_0 \frac{a^2}{2M_Y^u} = 4.6$; at about $\sigma_Y^u < 2\sigma_Y^i$ the non-shakedown collapse mode is incremental (otherwise alternating plasticity prevails). That non-shakedown collapse curve, under which the structure is safe, is presented in the plane of dimensionless external frequency ($K = k_U a$) and load ($Q = q_1^U \frac{a^2}{2M_Y^u}$) parameters in Fig. 5. As K approaches the natural frequency of the structure, shakedown limit on the fluctuating load amplitude Q decreases drastically.

Initial and Ultimate Yield Stresses

The initial yield stress σ_Y^i defining the cyclic or alternating plasticity collapse mode and the ultimate yield stress σ_Y^u determining the incremental collapse one are two basic plastic parameters for shakedown safety assessment of elastic plastic structures under variable and cyclic loads (Pham 2008, 2010).

Plastic deformation often starts from microscopic through mesoscopic to macroscopic scales without clear boundary. For high-cycle processes, σ_Y^i should be taken as small as the fatigue limit, since it defines that mode of collapse. For other ranges of cycles, σ_Y^i can be given higher values according to the respective fatigue curve.

The ultimate yield stress σ_Y^u obtained in the standard monotonic loading experiment is often reached at the large plastic deformation (up to about 100%), which falls far outside the small deformation assumption of the classical plasticity theory and shakedown theorems. In addition, the design requirement of many structures

would not allow excessive global configuration changes due to the large plastic deformations. Hence σ_Y^u for our shakedown analysis can be taken as the yield stress corresponding to some allowable small amount of plastic deformation from the monotonic loading experiment, such as that from the broadly used Ramberg-Osgood empirical formula

$$\sigma_Y = K(\epsilon^p)^n, \quad (19)$$

where ϵ^p – the plastic deformation; K – strength coefficient, n – strain hardening exponent. The most prominent yield strength is $\sigma_Y^{(0.2)}$ corresponding to the amount 0.2% of plastic deformation – considered as the first significant amount of irreversible strain. Note that though a local plastic deformation at the amount 0.2% may be insignificant for the global geometry of a structure because of the global compatible strain constraint, when a global incremental mechanism is formed at $\sigma_Y^u = \sigma_Y^{(0.2)}$ – a significant configuration change of the structure is expected.

For cyclic softening materials in multicycle processes, σ_Y^u may be even smaller and should be taken from the respective multicycle loading experiments, instead of the monotonic loading ones.

Cross-References

- [Contact Elasto-Plasticity](#)
- [Cyclic Loading and Cyclic Stress](#)
- [Fatigue](#)
- [Fatigue Limit](#)
- [Fatigue Strength-Load Cycle Relationships for Ferrous Materials](#)

References

- D.A. Gokhfeld, O.F. Cherniavski, *Limit Analysis of Structures at Thermal Cycling* (North Holland, Alphen aan den Rijn, 1980)
- W.T. Koiter, General theorems for elastic-plastic solids, in *Progress in Solids Mechanics*, ed. by I.N. Sneddon, R. Hill (North-Holland, Amsterdam, 1963), p. 165
- A. König, *Shakedown of Elastic-Plastic Structures* (Elsevier, Amsterdam, 1987)
- D.C. Pham, Safety and collapse of elastic-plastic beams against dynamic loads. *Int. J. Mech. Sci.* **42**, 575 (2000)
- D.C. Pham, Shakedown theory for elastic-perfectly plastic bodies revisited. *Int. J. Mech. Sci.* **45**, 1011 (2003a)
- D.C. Pham, Plastic collapse of a circular plate under cyclic loads. *Int. J. Plasticity* **19**, 547 (2003b)
- D.C. Pham, Shakedown theory for elastic plastic kinematic hardening bodies. *Int. J. Plasticity* **23**, 1240 (2007)
- D.C. Pham, On shakedown theory for elastic-plastic materials and extensions. *J. Mech. Phys. Solid.* **56**, 1905 (2008)
- D.C. Pham, Shakedown working limits for circular shafts and helical springs subjected to dynamic fluctuating loads. *J. Mech. Mater. Struct.* **5**, 447 (2010)

- D.C. Pham, H. Stumpf, Kinematical approach to shakedown analysis of some structures. *Quarter. Appl. Math.* **52**, 707 (1994)
- T.N. Tran, G.R. Liu, H. Nguyen-Xuan, T. Nguyen-Thoi, An edge-based smoothed finite element method for primal-dual shakedown analysis of structures. *Int. J. Numeric. Method. Eng.* **82**, 917 (2010)
- D. Weichert, G. Maier (eds.), *Inelastic behaviour of structures under variable repeated loads* (Springer, Wien, 2002)

Shape Function in FEM

- [Finite Element Method for Fluid Film Bearings](#)

Shaper Cutters

- [Gear Cutting Tools](#)

Shaping

- [Gear Manufacturing Machines](#)

Shaving

- [Gear Surface Treatment](#)

Shaving, Broaching

- [Gear Manufacturing Machines](#)

Shear Dependence of Viscosity

SCOTT BAIR

School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Synonyms

[Non-Newtonian viscosity](#); [Pseudoplasticity](#); [Shear-thinning](#)

Definition

Shear viscosity, η , as opposed to extensional viscosity or bulk viscosity, is the ratio of a shear stress, τ , acting upon a plane divided by the gradient, $\dot{\gamma}$, normal to the plane of the component of velocity in the direction of the shear stress. The limiting low shear viscosity, μ , is the value of η that is approached as the stress or rate go to zero. The shear dependence of the generalized viscosity, η , is addressed in this article.

Scientific Fundamentals

Viscosity

The one property of the fluid lubricant that must always appear in the Reynolds equation is the shear viscosity (Cameron 1970), given in units of the product of stress and time. Elongational viscosity may be important in squeeze films and bulk viscosity may be important when a rapid change in pressure or temperature occurs, but these are not addressed here. The shear viscosity of the liquids that are used as lubricants decreases with increasing shear stress (or shear rate) after a sufficiently large stress (or rate) is attained. There is presently no simple theory that can accurately predict the shear dependence of the viscosity of typical lubricants, although non-equilibrium molecular dynamics simulations (Bair et al. 2002) have shown good agreement with viscometer measurements for a pure organic liquid, squalane. Viscosities extracted from tribological measurements have not proven to be accurate (Bair et al. 2000). See the article “► [High Pressure Viscometers](#)” for a description of pressurized Couette and capillary viscometers for viscosity measurement at sufficiently large shear stress to show shear dependence in all liquid lubricants.

Molecular Alignment

It was discovered early in the study of elastohydrodynamic lubrication (EHL) that a Newtonian description of the shear response, even with viscous heating, cannot explain the friction and film thickness that can be measured. One early attempt to explain the required shear dependence of viscosity involved viscoelasticity. If a single-mode Maxwell model is assumed, then the shear rate is the sum of an elastic strain rate and a viscous strain rate. If a particular form of objective time derivative of the stress is used, then the shear stress will be reduced from the Newtonian expectation when the shear stress becomes greater than the elastic modulus (Hutton 1973). The shear modulus of the liquid may be determined from ultrasound measurements, and it is too large to explain contact behavior.

A second early approach involved the measurement of friction in an EHL contact as a function of sliding velocity. The shear stress averaged over the contact area can then be plotted against the shear rate averaged across the film and over the contact area. This plot was then treated as a flow curve of shear stress versus shear rate for the conditions of the oil inlet temperature and the average contact pressure.

Neither of the above concepts proved reliable; they did not agree with viscometer measurements (Bair et al. 2000) and the properties that they produced are not useful for the calculation of both EHL film thickness and friction. The Einstein-Debye relation (Bair 2007) for the rotational relaxation time of a molecule is:

$$\lambda_{EB} = \frac{\mu M}{\rho R_g T} \quad (1)$$

where M is the molecular weight, ρ is the mass density, and R_g is the gas constant ($8,314 \text{ Pa}\cdot\text{m}^3\cdot\text{kmol}^{-1}\cdot\text{K}^{-1}$). Figure 1 shows the measured viscosity of a diester ($M = 391 \text{ kg/kmol}$) plotted versus shear rate. See the article on “► High Pressure Viscometers” for a description of the pressurized Couette viscometer. The curves plotted through the data represent the Carreau viscosity equation (Carreau 1972):

$$\eta = \mu_2 + (\mu - \mu_2) \left[1 + (\lambda \dot{\gamma})^2 \right]^{\frac{(n-1)}{2}} \quad (2)$$

with the value of the power-law index, $n = 0.38$, having been adjusted to fit the rate of change of viscosity with shear rate, $\dot{\gamma}$, in the power-law regime. The second Newtonian does not appear in Fig. 1, $\mu_2 = 0$. In (2), n defines the slope, $n - 1$ the descending (non-Newtonian) part of

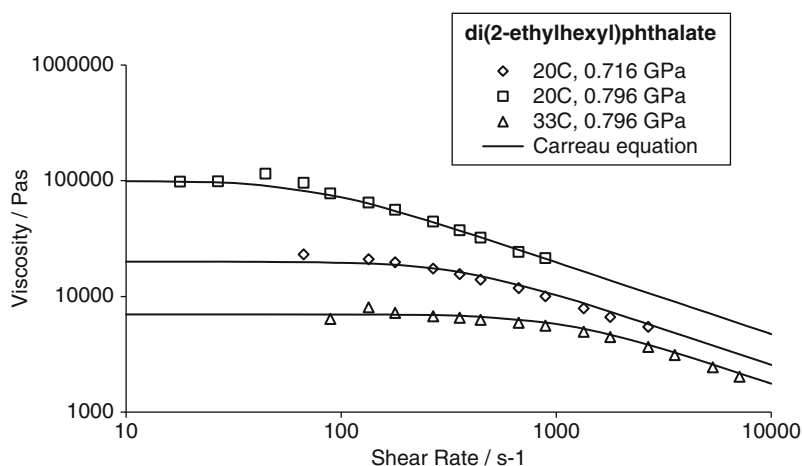
the flow curve, and $\lambda = \lambda_{EB}$ defines the beginning, at $\lambda \dot{\gamma} = 1$, of shear thinning. The theoretical rotational relaxation time is an excellent predictor of the onset of shear-thinning when the liquid is a single component of well-defined molecular weight. The relationship between shear-thinning and the rotation and elongation of the molecules of the liquid can be substantiated by observing the occurrence of birefringence as the non-linear shear response develops.

Measurement Techniques

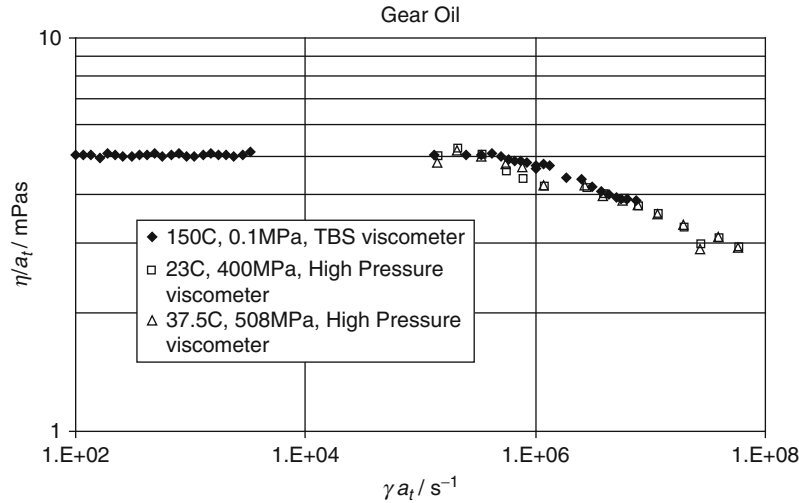
Commercial viscometers are available for measurement of shear dependent viscosity at ambient pressure (ASTM-D4683 1996). These thin-film Couette viscometers generate flow curves for high temperatures that can be successfully compared with the measurements by high-pressure Couette viscometers using the powerful technique of time-temperature-pressure superposition, discussed in the next section. See Fig. 2, where data from a TBS viscometer are compared with data from a high pressure viscometer shown in Fig. 5 of the article “► High Pressure Viscometers.”

Prior to the development of pressurized thin-film Couette viscometers, the capillary viscometer was the only source of shear dependent viscosity data at elevated pressure. Measurements of the viscosity of a polymer-blended mineral oil are shown in Fig. 3.

The curves fitted to the data represent the equation introduced by Bair and Khonsari 1996, similar to the Carreau equation (2) and the independent variable is now stress, τ .



Shear Dependence of Viscosity, Fig. 1 The shear dependence of viscosity of dioctyl phthalate, measured in a pressurized Couette viscometer (From Bair (2007) with permission of Elsevier)



Shear Dependence of Viscosity, Fig. 2 Flow curve generated with a TBS viscometer at ambient pressure compared with flow curves at elevated pressures (The high temperature, high shear data are courtesy of Frederic Jarnias of the Total Oil Company)

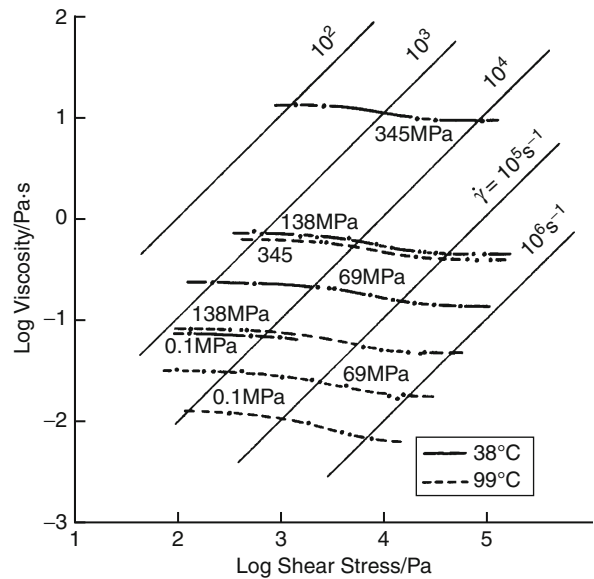
$$\eta = \mu_2 + (\mu - \mu_2) \left[1 + \left(\frac{\tau}{G} \right)^2 \right]^{\frac{(1-\lambda)}{2}} \quad (3)$$

In Fig. 3, the second Newtonian viscosity, μ_2 , is not zero and the power-law slope does not fully develop because the first and second Newtonian viscosities are not sufficiently different. These two representations of shear thinning, (2) and (3), can be related by writing the rotational relaxation time as $\lambda = \mu/G$. Therefore, for a liquid with a single, well-defined molecular weight, $G = \rho R_g T/M$. For the gear oil of Fig. 2, $G = 3.8\text{ kPa}$ and $n = 0.88$ and $\mu_2 = 0$.

Extreme care must be taken when viscometer measurements are interpreted as constitutive behavior. There is a long history of temperature variations in viscosity brought about by viscous heating in the viscometer having been mistaken for shear dependency (Hersey and Zimmer 1937). Also, there is a complication in the measurement of viscosity at ambient pressure that does not occur at high pressures. When the shear stress of a measurement approaches a value equal to the ambient pressure, normal stress differences can contribute to the tensile stress in the liquid in the principal tensile direction, resulting in cavitation. The subsequent reduction in viscosity has been mistaken for shear-thinning as well (Bair 2007).

Time-Temperature-Pressure Superposition

If a measurement of viscosity as a function of shear rate (stress) is made at an experimentally convenient reference temperature, T_R , and pressure, p_R , the resulting flow curve



Shear Dependence of Viscosity, Fig. 3 Flow curves generated with a high-pressure capillary viscometer for a mineral oil blended with 4% polyalkylmethacrylate (Adapted from Novak and Winer (1968))

can be generalized to another state, if $\mu(T, p)$ and $\rho(T, p)$ are known, by invoking (1).

$$\lambda = \lambda_R \frac{\mu_R T_R}{\mu_R \rho T} \quad \text{or} \quad G = G_R \frac{\rho T}{\rho_R T_R} \quad (4)$$

The product ρT varies slowly with temperature and pressure in comparison to μ and, therefore, $G = G_R$ or $\lambda = \lambda_R \mu / \mu_R$ are approximations sufficient for many tribological calculations. The horizontal shift factor $a_t = \mu / \mu_R$ comes from this approximation, which can be seen in Fig. 2 to provide a master curve for the reference state of 150°C at atmospheric pressure.

Two-Dimensional Flow

If the lubricated contact cannot be modeled as a line contact, then the viscosity function must be slightly different from (2) and (3). The shear rate, $\dot{\gamma}$, in (2) is replaced by the equivalent shear rate, $\dot{\gamma}^*$, or the shear stress, τ , in (3) is replaced by the equivalent shear stress, τ^* . If the cross-film direction is the z -direction and the velocities in the normal directions are u and v , then:

$$\dot{\gamma}^* = \sqrt{\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2} \quad (5)$$

Similarly, if the cross-film direction is the z -direction, then:

$$\tau^* = \sqrt{\sigma_{xz}^2 + \sigma_{yz}^2} \quad (6)$$

where σ_{xz} and σ_{yz} are the shear stresses in the normal directions.

Mixtures

If a lubricant is a mixture of components, each providing a substantial contribution to the viscosity, the shear dependence becomes more complicated. A qualitative

assessment of the behavior can be had from application of the idealized Arrhenius mixing rule:

$$\ln[\eta(\tau)] = \sum_{i=1}^N c_i \ln[\eta_i(\tau)], \quad \sum_{i=1}^N c_i = 1 \quad (7)$$

for concentrations c_i . A quantitative description of the shear dependence must come from a viscometer measurement of the mixture viscosity. In that case, a more accurate model is the Ree-Eyring (Kim et al. 1960) equation:

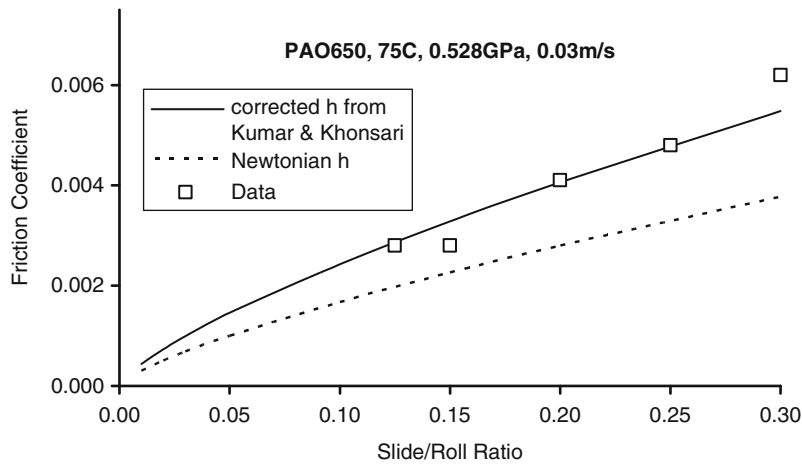
$$\eta = \mu \sum_{i=1}^N \frac{f_i}{\lambda_i \dot{\gamma}} \sinh^{-1}(\lambda_i \dot{\gamma}), \quad \sum_{i=1}^N f_i = 1, \quad N > 1 \quad (8)$$

when it is fitted to the data. Because of the approximation of power-law behavior with \sinh^{-1} terms, the Ree-Eyring equation (8) requires a large number, N , of \sinh^{-1} terms to describe the mixture of only two components, sometimes as many as five (Kim et al. 1960), and the characteristic times, λ_i , are not necessarily the rotational relaxation times.

Key Applications

EHL Film Thickness

EHL film thickness may be accurately calculated from the classical Newtonian formulas for $h_{\text{Newtonian}}$ only when the liquid is substantially Newtonian (and isothermal) throughout the inlet zone. Otherwise, either a full numerical simulation must be performed with the shear dependent viscosity or, if the rheology is simple (a single component liquid without a second Newtonian) one of several correction formulas may be applied. One such formula for circular EHL



Shear Dependence of Viscosity, Fig. 4 Friction measured in a circular contact and calculated using a classical Newtonian film thickness formula and, alternatively, applying the shear-thinning correction of (9)

contact with rolling velocity, V_r , is from Kumar and Khonsari (Kumar and Khonsari 2008).

$$h_c = h_{\text{Newtonian}} \left(1 + 1.75 \frac{\mu_0 V_r}{G h_{\text{Newtonian}}} \right)^{-1.16(1-n)^{1.23}} \quad (9)$$

Friction

For low-pressure contacts at high temperature, a reasonable calculation of the isothermal friction may be obtained from integration of the local shear stress, $\eta(p, \dot{\gamma})\dot{\gamma}$, over the contact area. It can be assumed for the calculation that the pressure distribution is Hertzian and that the shear rate is uniformly equal to $(V_2 - V_1)/h_c$.

This simple approach is successfully demonstrated in Fig. 4, where the film thickness correction (9) has been applied. However, as shown in the figure, an accurate statement of the generation of the central film thickness is essential to an understanding of the friction. The use of an effective viscosity to arrive at the correct film thickness would have rendered the friction calculation impossible without further adjusting of viscosity. The simple calculation must fail at higher sliding when the flow is non-isothermal or at higher pressure when the shear banding limit to the shear stress is reached.

Full EHL Simulation

Another approach to numerical simulation has emerged wherein the viscosity is not treated as an adjustable parameter. The viscosity obtained from viscometers has been utilized in EHL simulations (Liu et al. 2007) with the result that the liquid properties responsible for the generation of a film and the dissipation of energy have been enumerated. This exciting approach is certain to result in many discoveries in the next few years.

Cross-References

- EHL Film Thickness Behavior
- EHL Governing Equations
- Elastohydrodynamic Lubrication
- Friction/Traction Behavior of EHL
- Hydrodynamic Lubrication
- Lubricant Non-Newtonian Effect on EHL
- Lubricant Viscosity
- Lubrication with a Non-Newtonian Fluid
- Newton's Law of Viscosity, Newtonian and Non-Newtonian Fluids

References

American Society for Testing Materials, *D4683-standard test method for measuring viscosity at high temperature and high shear rate by tapered bearing simulator* (1996), pp. 1–6

- S. Bair, On the concentrated contact as a viscometer, in *Proceedings of IMechanical Engineering, Part J: Journal of Engineering Tribology*, vol. 214 (2000), pp. 515–521
- S. Bair, *High Pressure Rheology for Quantitative Elastohydrodynamics* (Elsevier Science, Amsterdam, 2007), pp. 168–169, 141–143
- S. Bair, M. Khonsari, An EHD inlet zone analysis incorporating the second Newtonian. *ASME J. Tribol.* **118**, 341–343 (1996)
- S. Bair, C. McCabe, P.T. Cummings, Comparison of non-equilibrium molecular dynamics with experimental measurements in the nonlinear shear-thinning regime. *Phys. Rev. Lett.* **88**(5), 058302-1–058302-4 (2002)
- A. Cameron, *Basic Lubrication Theory* (Longman, London, 1970), pp. 1–29
- P.J. Carreau, Rheological equations from molecular network theories. *Trans. Soc. Rheol.* **16**(1), 99–127 (1972)
- M.D. Hersey, J.C. Zimmer, Heat effects in capillary flow at high rates of shear. *J. Appl. Phys.* **8**, 359–363 (1937)
- J.E. Hutton, Theory of rheology, in *Interdisciplinary Approach to Liquid Lubricant Technology*, ed. by P.M. Ku (NASA, Washington, 1973), pp. 187–261
- W.K. Kim, N. Hirai, T. Ree, H. Eyring, Theory of non-Newtonian flow III. A method for analyzing non-Newtonian flow curves. *J. Chem. Phys.* **31**(2), 358–361 (1960)
- P. Kumar, M. Khonsari, EHL circular contact film thickness correction factor for shear-thinning fluids. *ASME J. Tribol.* **130**, 041506 (2008)
- Y. Liu, Q.J. Wang, S. Bair, P. Vergne, A quantitative solution for the full shear-thinning EHL point contact problem including traction. *Tribol. Lett.* **28**, 171–181 (2007)
- J.D. Novak, W.O. Winer, Some measurements of high pressure lubricant rheology. *ASME J. Lubr. Technol.* **F 90**(3), 580–591 (1968)

Shear Localization, the Limiting Stress, and Other Forms of Liquid Failure

SCOTT BAIR

School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Synonyms

Fracture; Slip

Definition

A rheological constitutive equation relates the stress with the deformation of a liquid. The velocity within a lubrication flow does not always conform to the prediction of an otherwise correct constitutive equation. Hutton lists several forms of liquid failure (Hutton 1973) in which the shear of a liquid results in material changes of which some are permanent (for example, molecular degradation) and some are temporary (for example, fracture). The forms of liquid failure that are discussed here are fracture in tension, cavitation, and cohesive slip.

Scientific Fundamentals

Fracture in Tension and Cavitation

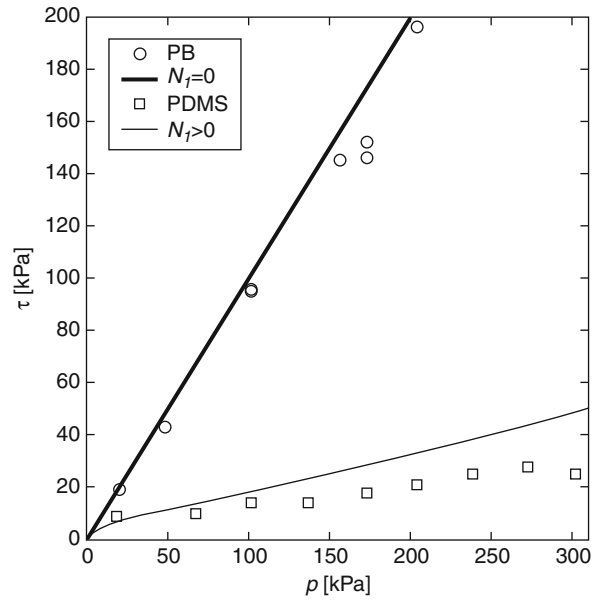
Cavitation is a form of failure that is extremely important to hydrodynamic lubrication. There are certain geometries that must have cavitation to generate load support. The usual strategy for hydrodynamic problems is to assume that cavitation is the result of vapor or gas coming out of the liquid when the pressure decreases to the vapor pressure or when $p = 0$. However, for carefully controlled experiments large negative pressure or hydrostatic tension may be metastable for liquids (Hiro et al. 1993; Fisher 1948). Cavitation seems to require more than simply a low pressure. In fact, cavitation may occur when only one principal stress becomes tensile while the mean mechanical pressure, p , is large and positive. This shear cavitation can be enhanced by the appearance of a large first normal stress difference, N_1 , in high rate shear (Kottke et al. 2005), as shown in Fig. 1, and is a hindrance in the measurement of viscosity at high shear rate at atmospheric pressure. Photographs of shear cavitation voids can be found in the reference. The criterion for cavitation can be written as:

$$\frac{N_1 + \sqrt{N_1^2 + 4\tau^2}}{2} > p \quad (1)$$

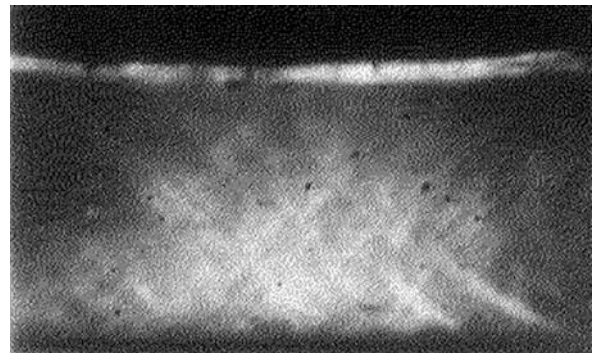
Slip Within the Liquid

It has been an unusual trait of friction in full-film elastohydrodynamic lubrication that the friction coefficient should never exceed about 0.015, and typically the limit is much lower. It seems as though a liquid possesses an internal friction coefficient. For a viscous liquid, the shear stress must ordinarily be determined by the product of shear rate and viscosity with allowance for the thermal softening that results from viscous heating. The viscosity is usually dependent upon shear rate and follows a power-law rule for stress greater than a critical value. See the article “► [Shear Dependence of Viscosity](#).” This description of the constitutive behavior of a liquid does not yield the broad plateau of rate-independence that is often observed in full-film friction measurements at high pressure.

Flow visualization experiments at high pressure have elucidated the cause of the rate independent friction in simple shear in full viscous films. Figure 2 is a micrograph obtained in a different type of high-pressure optical cell (Bair 2007) in which an ambient pressure of 280 MPa was maintained while the oil was squeezed between a 1-mm diameter steel pin and a hardened steel flat. The flow localizes along shear bands. Bands operate for a short time, disappear, and new bands form.



Shear Localization, the Limiting Stress, and Other Forms of Liquid Failure, Fig. 1 The shear stress at first appearance of cavitation in simple shear as a function of pressure for polybutene, PB, and polydimethyl siloxane, PDMS. The first normal stress difference, N_1 , reduces the stress required for cavitation. The curves are (1) (Reproduced from Kottke et al. (2005))



Shear Localization, the Limiting Stress, and Other Forms of Liquid Failure, Fig. 2 A micrograph of a squeeze film at high pressure showing the two orientations of shear bands

The Mohr-Coulomb theory correctly predicts the orientation of the two types of bands shown in Fig. 2 when applied to simple shear. This theory predicts slip along surfaces for which the ratio of shear stress to compressive

normal stress is equal to the internal friction coefficient of the liquid (Bair 2007).

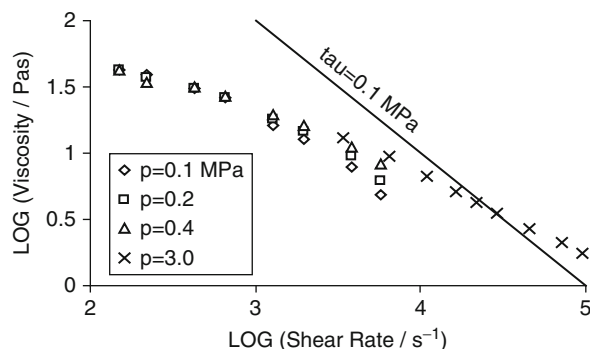
Key Applications

EHL Friction

For most elastohydrodynamic conditions, the calculation of full-film friction will require a limiting stress assumption. The Mohr-Coulomb theory provides the motivation for limiting the local shear stress to a fraction of the local pressure. This fraction is a characteristic property of the liquid that is slightly temperature dependent.

High Shear Viscometry

The shear dependence of viscosity is essential for understanding the film formation and friction behavior of hydrodynamic lubrication films. Measurements performed at atmospheric pressure must be limited to shear stress less than that calculated from the failure criterion of (1). Figure 3 illustrates the problem. Measurements at pressure near atmospheric for silicone oils often show a power-law exponent, n , that is very small and flow curves for different molecular weights and temperatures converging along a curve representing shear stress of approximately 0.1 MPa, the pressure of the atmosphere. In Fig. 3, it can be seen that repeating the measurement at slightly elevated pressure brings the power-law exponent back to a more reasonable value of $n = 0.4$ and allows measurement at stress, τ , exceeding 0.1 MPa.



Shear Localization, the Limiting Stress, and Other Forms of Liquid Failure, Fig. 3 Flow curves generated in a Couette viscometer for a silicone oil at various pressures showing shear cavitation (Reproduced from Bair (2007) with permission of Elsevier)

Cross-References

- Cavitation Formation and Modeling
- Cavitation Phenomena and Numerical Analysis
- EHL Governing Equations
- Elastohydrodynamic Lubrication
- Elastohydrodynamic Lubrication (EHL)
- Friction/Traction Behavior of EHL
- Gear Lubrication
- High Pressure Viscometers
- Hydrodynamic Lubrication
- Lubricant Non-Newtonian Effect on EHL
- Lubricant Viscosity
- Lubrication with a Grease
- Lubrication with a Non-Newtonian Fluid
- Newton's Law of Viscosity, Newtonian and Non-Newtonian Fluids
- Rheological Measurement Methods and Equipment
- Shear Dependence of Viscosity
- Stress-Induced Lubricant Degradation and Viscosity Loss
- Temperature and Pressure Dependence of Viscosity
- Thermal Effect on EHL
- Thermal EHL Theory

References

- S. Bair, *High Pressure Rheology for Quantitative Elastohydrodynamics* (Elsevier Science, Amsterdam, 2007), pp. 199–200
- J.C. Fisher, The fracture of liquids. *J. Appl. Phys.* **19**, 1062–1067 (1948)
- K. Hiro, K. Nishii, Y. Ohde, Y. Tanzawa, Raising of negative pressure to around -200 bar for some organic liquids in a metal Berthelot tube. *J. Phys. D Appl. Phys.* **26**(8), 1188–1191 (1993)
- J.E. Hutton, Theory of rheology, in *Interdisciplinary Approach to Liquid Lubricant Technology*, ed. by P.M. Ku (NASA, Washington, DC, 1973), pp. 187–261
- P.A. Kottke, S.S. Bair, W.O. Winer, Cavitation in creeping shear flows. *AIChE J.* **51**(8), 2150–2170 (2005)

Shear Thinning

- Newton's Law of Viscosity, Newtonian and Non-Newtonian Fluids

Shear Transformation Zone (STZ)

- Crack Growth in Noncrystalline Solids

Shear-Thinning

- Shear Dependence of Viscosity

Shoe-Surface Friction

- [Tribology in Daily Life: Footwear-Surface Interactions in Pedestrian Slips](#)

Short Cracks

- [Growth Characteristics of Small Fatigue Cracks](#)

Short-Range Order (SRO)

- [Crack Growth in Noncrystalline Solids](#)

Shot Peening

MAYURAM M. M.

Department of Mechanical Engineering, Machine Design Section, Indian Institute of Technology Madras, Chennai, TN, India

Definition

Shot peening is a cold working process wherein a stream of shots are propelled at a high velocity onto the surface to be treated under controlled conditions. The process is applied on engineering components to produce a compressive residual stress on a layer of material in the surface and subsurface regions in order to enhance resistance to metal fatigue and some forms of stress corrosion.

Scientific Fundamentals

Introduction

Shot peening is a widely used cold working process, which produces a compressive residual stress on a layer of material in the surface and subsurface regions. The residual stresses are induced when the surface is bombarded by small, spherical media called *shots*, to a sufficient intensity under controlled conditions. The process, mechanism, process parameters, their effects and selection, and the benefits and applications of the process are discussed here.

The Process

During the shot peening process, a stream of shots is propelled at a high velocity onto the surface to be treated under controlled conditions. Each piece of shot that strikes the material acts as a tiny peening hammer, creating a small indentation or dimple on the surface. Sufficient energy is needed to create the dimple, as the surface of the material must yield in tension. As the dimples overlap with random impacts, the entire surface is effectively elongated, driving the surface layer of deformed material into compression.

The Mechanism

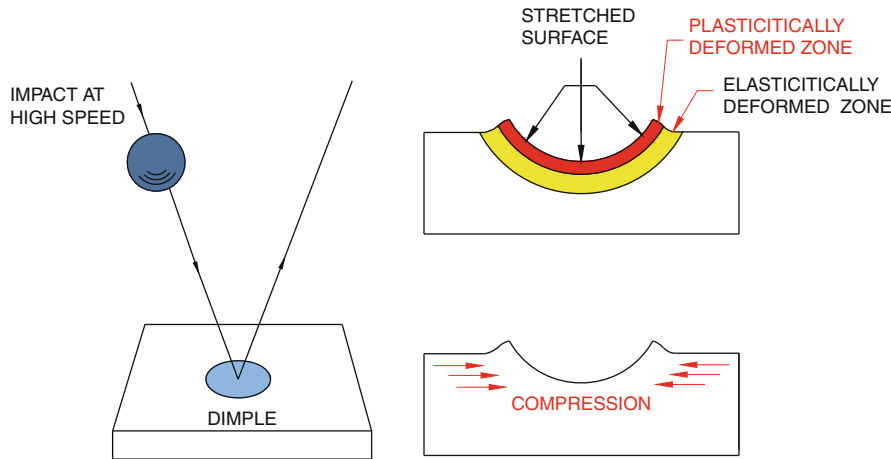
The compressive layer is formed by a combination of surface and subsurface compression developed by Hertzian loading combined with lateral displacement of the surface material around each of the formed dimples. Sufficient intensity causes plastic flow of surface metal at the instant of contact, stretching it radially (Fig. 1). The metal beneath this layer is stressed but not plastically deformed. Thus, the large volume of material below the surface, which is in an elastic state, tries to recover or regain its original shape, thereby producing, below the dimple, a hemisphere of cold-worked material highly stressed in compression. In the stress distribution that results, the surface metal has induced residual compressive stress parallel to the surface, while metal beneath has reaction-induced tensile stress. The surface compressive stress is several times greater than the subsurface tensile stress.

The effect usually extended to about 0.100–0.250 mm (0.004–0.010 in.), but may extend as much as 0.50 mm (0.02 in.) below the surface under some favorable conditions.

Effectiveness – Surface Coverage and Peening Intensity

The effectiveness of the process depends on two principal parameters, namely the surface coverage and the peening intensity. The randomly striking shots must uniformly spread the effect all over the surface; *surface coverage* is the term used to describe this aspect. The parameter associated with the depth and degree to which the effect is imparted is the peening intensity.

Surface coverage is a measure of how completely an area has been hit by the myriad of impinging shot particles. Without 100% coverage or saturation, the improvement in fatigue characteristics or the full benefit produced by shot peening cannot be realized. Surface coverage directly depends upon the exposure time, time to which a specific area is exposed to the impinging shots. As the exposure time increases coverage also increases by a nonlinear function; 100% coverage is the theoretical limit and difficult to obtain.



Shot Peening, Fig. 1 Mechanism of residual stress creation due to dimple formation

Direct methods for measuring coverage are seldom used and desired coverage for saturation is associated with exposure time through a measure of change in intensity, as explained later.

The relative work done to the surface is called the *peening intensity*. The depth up to which the effect can extend and the magnitude of residual stress that can be induced are functions of the peening intensity and the characteristics of the material treated: the higher the intensity, the higher the effectiveness.

Because of the difficulty in quantitatively measuring the actual coverage and peening intensity directly, a simpler comparative method has been devised to measure these parameters.

If a flat strip of metal is shot peened on one side only it will curl slightly and produce a convex surface from the side that has been peened; the degree of curvature is a measure of the peening intensity, with the strip curling more at higher intensities. For reliability, reproducibility, and comparison, measurements are done using a standard strip and this test is popularly known as the Almen test, after the man who formalized this method.

Almen Test- Strip, Holder, and Gauge

The SAE standard J442 describes the test strip, strip holder, and gauge used in measuring shot peening intensity. The standard strip is called an *Almen strip*. It is made from spring steel of carefully controlled quality to a size within close tolerances. The strips are made out of cold-rolled, SAE 1070 spring steel, and have a specified hardness of 44–50 HRC. It is available in three thicknesses, C, A, and N; the C strip is thickest and N strip is thinnest. Arc height is the measure of the curvature of a test strip that has been

peened on one side only. The strip is placed and retained magnetically or with the aid of a flat spring against two pairs of ball contacts, a fixed distance apart. The curvature or arc height of the strip is measured with the aid of a dial gauge. The gauge is zeroed with the unpeened strip in position. After peening, the strip is replaced against the contacts with the unpeened side towards the dial gauge stem and the Almen arc height is read directly in millimeters or thousandths of an inch.

The three different strip thicknesses cater to different extremes of peening intensity. For most applications, an A strip would be used, and if this give a deflection after peening of 0.4 mm (0.016 in.) the intensity would be expressed as 0.4 mm A (0.016 in. A). An N strip would be used for lighter peening, giving less than 0.15 mm A (0.006 in. A). The C strip is used for heavy peening, having an intensity greater than 0.55 mm A (0.22 in. A).

Generally, arc height N is three times arc height A and a C reading is about 0.3 of that with an A strip. In practice, 80% of all peening requirements lie between 0.3 mm (0.012 in.) A and 0.5 mm (0.020 in.) A.

Peening Intensity

The depth of compressed layer to be produced by peening is a factor in selecting peening intensity. Peening intensity is mainly governed by the velocity, hardness, size, and weight or throughput of the shot or pellets, and by the angle at which the stream of shot impinges against the surface of the workpiece. Arc height at or above saturation of the Almen strip is the standard measure of the effectiveness of the peening operation on a specific part. Intensity is usually expressed as the arc height of an Almen test strip, at or more than saturation coverage.

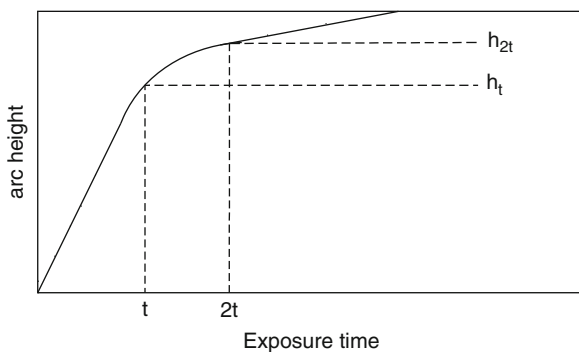
Surface Coverage-Exposure Time and Saturation

Saturation is associated with the change in intensity variation with exposure time. To identify it, a curve is plotted with the arc height measured and the corresponding exposure time using a number of experimentally determined points. The earliest point at which a change in arc height of 10% or less prevails when doubling the exposure time is the saturation point (Fig. 2). The coverage at this exposure is the saturation coverage. Peening at less than saturation is ineffective and often the best possible benefits are achieved by increased coverage, which is usually expressed as a multiple of exposure time required to produce saturation. For example, two times the saturation exposure time used is denoted as 200% coverage, implying that each point on the surface was impacted at least two or more times.

The shot peening process is defined by parameters that include the size and type of shot used, the Almen intensity achieved, and the coverage needed.

Equipment

The equipment used in shot peening is essentially the same as that used in abrasive blast cleaning, except for certain auxiliary equipment made necessary by the more stringent controls imposed in the shot peening process. The principal components of shot peening equipment are a shot-propelling device, which govern the shot velocity and the angle of impingement, shot-cycling arrangements, and a work-handling device that controls the coverage. The entire process is carried out in a closed enclosure for safety as well as to facilitate shot recycling. All portions of equipment that are exposed to the stream of shot are enclosed to confine the shot and permit it to be recycled.



Shot Peening, Fig. 2 Variation of peening intensity (arc height) with exposure time. Coverage is considered full (100%) at time t if doubling exposure to time $2t$ results a change in arc height less than 10%

Propulsion of Shot

Two methods of propulsion of the shot are used widely in shot peening. One uses a motor-driven bladed wheel, rotating at high speed and the other uses a continuous stream of compressed air and entrains the shot along the flowing stream of air.

In the wheel method, shot is propelled by a bladed wheel that uses a combination of radial and tangential forces to impart the necessary peening velocity to the shot. The position on the wheel from which the shot is projected is controlled to concentrate the peening blast in the desired direction. Among the advantages of the wheel method of propulsion are easy control of shot velocity when equipped with a variable speed drive, high production capacity, lower power consumption, and freedom from the moisture problem encountered with compressed air.

The air blast method introduces the shot, either by gravity or by direct pressure, into a stream of compressed air directed through a nozzle onto the work to be peened. Aside from being more economical for limited production quantities, the air blast method can develop higher intensities with small shot sizes, permits the peening of deep holes and cavities by using a long nozzle, consumes less shot in peening small areas on intricate parts, and has lower initial cost, especially when a source of compressed air is already available.

Cycling of Shot

Equipment for shot recycling consists of devices for the separation and removal of dust and undersize shot from the used shot mix. The shot will deform or fracture during use, leaving broken pieces that can cause damage in the form of sharp notches upon impact with the surface. The media must therefore be constantly screened to remove broken shot and dust. In better quality shot blast equipment this is achieved by an efficient air wash system that removes undersize particles by passing a controlled air-stream upon the falling stream of peening media that has been precipitated by centrifugal means in a cyclone. The effectiveness of the separator depends on careful control of the velocity of the air. For final elimination of oversize unwanted debris, the reusable particles pass through a vibrating sieve. Shot-adding devices automatically replenish to maintain an adequate quantity of shot in the machine at all times. They are equipped with a capacitance switch or similar device to control the level of shot in the storage hopper and to add shot, as required, from a supply hopper.

Selection of Intensity

The lowest peening intensity capable of producing the desired compressive stress is the most efficient and least

costly, because the peening process can be achieved with the minimum exposure time and lowest shot handling.

Selection of Shot and Shot Size

The “media” or the shot used for peening is made of iron, steel, or glass and can be metallic or ceramic. Metallic shot is mostly made of either cast steel or from cut wire blasted against a carbide plate to form a nearly spherical shape. Cut wire shot can be manufactured from virtually any alloy to avoid elemental contamination. Ceramic shot is typically zirconium oxide or glass bead. Metallic shot is designated by numbers according to size; it ranges from S70 to S930. The shot number is approximately the same as the nominal diameter of the individual pellets in ten thousandths of an inch.

Cast steel shot is the most widely used peening medium. With suitable heat treatment, cast steel shot has a useful life many times that of cast iron shot. To maximize the peened effect, shot should always be as hard as the workpiece. For most applications 90% of the hardness measurements made on a representative sample should fall within the range equivalent to 40–50 HRC. For hard metals, special hard cast steel shot, 57–62 HRC, should be used. Its improved impact and fatigue properties markedly lower the rate of shot breakage, increase peening quality, and extend the life of components peening machines.

Cast iron shot or chilled iron is brittle, with an as-cast hardness of 58–65 HRC. Its initial cost is lower and its inherently high hardness yields higher peening intensities for a given shot size, in comparison to softer materials. A high rate of shot breakage complicates the control and increases the cost of equipment maintenance and cost of shot, because broken shot must be eliminated for best results.

Where ferrous contamination is a problem, nonmetallic, glass, or ceramic shots are used. Glass beads are used for peening precision aircraft components of stainless steel, titanium, aluminum, magnesium, and other metals that might be contaminated by iron or steel shot. They are also used for peening thin sections. Relatively low shot peening intensities, seldom exceeding 0.15–0.25 mm (0.006–0.010 in.) Almen A, can only be achieved. Glass beads can be used in either wet or dry peening processes.

Shot Size

Selection of an appropriate shot size is critical. For maintaining or improving surface finish larger size shots are recommended; higher intensity can also be achieved, but throughput (volume of shot thrown per unit time) will be lower. Apart from considerations of peening intensity, which are directly affected by shot size, the size of shot should be small enough to fit the smallest

inside radius or fillet being peened, preferably less than one third of the smallest fillet radius used. In compressed air-type direct blast systems, the size should suit the nozzle through which the shot is propelled such that continuous flow without block or interruption is maintained.

Benefits

Cold working can be an advantage in a number of applications where the inherent increase in yield and ultimate strengths enhance fatigue performance along with the layer of surface compression. Nearly all fatigue and stress corrosion failures originate at the surface of a part, but cracks will not initiate or propagate in a compressively stressed zone. Thus the compressive stresses induced through shot peening confer resistance to metal fatigue and to some forms of stress corrosion. The tensile stresses deep in the part are not as problematic as tensile stresses on the surface because cracks are less likely to start in the interior. In most modes of long-term failure, the common denominator is tensile stress. Tensile stresses attempt to stretch or pull the surface apart and may eventually lead to crack initiation. Because crack growth is slowed significantly in a compressive layer, increasing the depth of this layer increases crack resistance. Shot peening is the most economical and practical method of ensuring surface residual compressive stresses.

A study done through the SAE Fatigue Design and Evaluation Committee showed the benefits of shot peening welds by comparing them with welds that didn't undergo this operation. The study claimed that the regular welds that would fail after 250,000 cycles can last 2.5 million cycles, and failures lie mostly outside the weld area. This is part of the reason that shot peening is a popular operation with aerospace parts. However, the beneficial pre-stresses can anneal out at higher temperatures.

Key Applications

Shot peening may be used for cosmetic effect. The surface roughness resulting from the overlapping dimples causes light to scatter upon reflection. Because peening typically produces larger surface features than sand-blasting, the resulting effect is more pronounced.

Shot peening also can induce the aerodynamic curvature in metallic wing skins used in advanced aircraft designs. Additional applications for shot peening include work hardening through cold work to improve wear characteristics, closing of porosity, improving resistance to intergranular corrosion, straightening of distorted parts, surface texturing and testing the bond strength of coatings, railway leaf springs, automobile leaf springs, helical

springs of all types, gears of all types, axle bearings, crankshafts, pneumatic drills, milling cutters, connecting rods, cylinder blocks, valve springs washers, and so on.

There are some problems associated with shot peening. The beneficial pre-stresses can anneal out at higher temperatures. In some cases the induced stresses gradually or rapidly disappear, a phenomenon known as *fading* which is mainly due to thermal relaxation, or hysteresis losses during cyclic loading; soft materials are more susceptible to such things. Another drawback is that, by its very nature, shot peening relies upon random impacts of the shot. Therefore, in order to achieve the coverage requirements, some regions will receive numerous impacts before adjacent areas are impacted at all. The result is a nonuniform and generally very highly cold-worked surface. Cold work levels range from 40% to more than 100% during creation of the layer of surface compression. In the work hardening materials, such as titanium and nickel alloys, peening-induced cold working can exhaust the ductility of the material, leaving a brittle surface layer. Shot peening damage in the form of “laps and folds” creates stress concentrations that reduce fatigue performance.

Cross-References

- [Fatigue](#)
- [Gear Surface Treatment](#)
- [Shakedown](#)

References

- ASM Committee on Shot Peening, Shot peening, in *Metals Handbook*, Vol 5, 9th edn. pp138–149
- K.J. Marsh (ed.), *Shot Peening – Techniques and Applications* (Shot EMAS Publishing, Warley, 1993)
- Manual on Shot Peening*, (SAE International, Warrendale, 2001). ISBN 978-0-7680-0868-5
- SAE J2277 Shot Peening Coverage; SAE International Jan 03
- SAE J2441 ‘Shot Peening’ SAE International Aug 2000
- SAE J443 (R) Procedures for Using Standard Shot Peening Test Strip SAE International Jan-2003

Shudder

- [Friction Modifiers](#)

SIE – Singular Integral Equations

- [Fretting Contact and Simulation](#)

SIF

- [Stress Intensity Factors](#)

SIF – Spherical Inverse Filter Method for Joint Simulation

- [Biotribological Joint Simulation System](#)

SIF – Spherical Inverse Filter Method for Spherical-Bearing Friction Prediction

- [Friction Prediction for Spherical Bearings](#)

SIF – Spherical Inverse Filter Method for Spherical-Bearing Wear Modeling

- [Wear Modeling of Spherical Bearings](#)

SIFM – Smooth Spherical Inverse Filter Method

- [Lubrication Theory for Spherical Bearings](#)

Silicon Carbide

- [Materials for Mechanical Seals](#)

Silicon Nitride

- [Materials for Mechanical Seals](#)

Siliconizing

DALIBOR VOJTĚCH

Department of Metals and Corrosion Engineering,
Institute of Chemical Technology, Prague, Czech Republic

Definition

Siliconizing is surface alloying of metallic materials with silicon to improve resistance against high-temperature oxidation, hot corrosion, carburization, and wear.

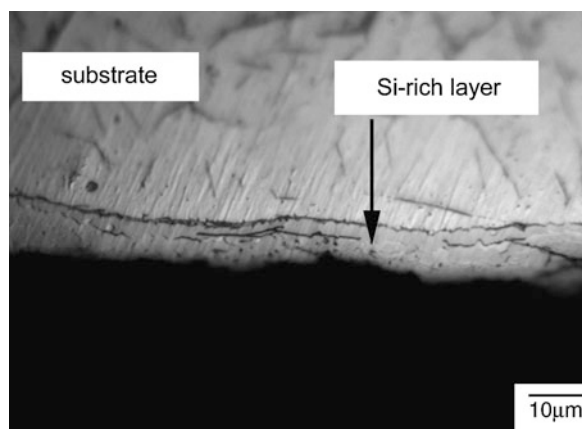
Scientific Fundamentals

Silicon has been shown to positively influence resistance of various materials, including steels, Ni-based alloys, Ti-based alloys, and others, against high-temperature oxidation, hot corrosion in salt melts, carburization, and wear. However, silicon additions to the bulk of alloys significantly modify their mechanical properties. Sufficient amounts of silicon result in the formation of hard and brittle silicides in the microstructure of alloys, which have undesirable effects on plasticity and toughness. For this reason, it appears more suitable to add silicon only to a surface layer of a material without influencing the bulk properties.

Methods

Siliconizing can be realized by various methods, for example, by pack cementation (sometimes termed *powder siliconizing*), chemical vapor deposition, physical vapor deposition, liquid phase siliconizing, laser surface alloying, or ion implantation. These methods can produce various siliconized layers with a thickness ranging between several micrometers and 1 mm.

Pack cementation is a widely used, very efficient, and inexpensive method to modify the surface of alloys. It consists simply of embedding a material to be coated in a powder and heating it at an appropriate temperature (around 1,000°C) for several hours under an inert or reducing atmosphere. Powder mixtures of silicon, inert fillers (Al_2O_3 , SiO_2), and halide salt activators (e.g., NaCl, NaF, NH_4Cl , etc.) are commonly used as siliconizing media (Bianco et al. 1991). A powder composed of 30 wt% Si, 67 wt% Al_2O_3 , and 3 wt% NaF is an example (Huang et al. 2006). During siliconizing, the activators react with silicon to form gaseous silicon halides, which then diffuse through the gaseous phase of the porous pack towards the material surface, followed by decomposition and deposition of silicon on the surface. Silicon then diffuses into a surface layer of material and its penetration



Siliconizing, Fig. 1 SEM image of a siliconized layer on titanium prepared by pack cementation at 1,100°C/3 h

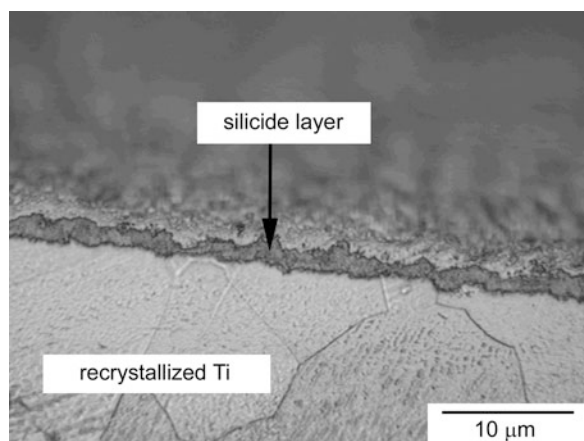
depth varies considerably with siliconizing temperature and time, chemical composition of a pack, and substrate. Figure 1 presents a cross-section of a siliconized layer on titanium prepared by pack cementation in pure Si powder at 1,100°C/3 h.

In chemical vapor deposition (CVD), a material to be coated is placed in a stream of a silicon-containing gas that is a diluted mixture of a volatile silicon compound, for example, SiCl_4 , SiCl_2H_2 , SiH_4 , or Si_2Cl_6 , and hydrogen (Klam et al. 1991). Concentrations of Si compounds in gas are generally around 1%. At a sufficiently high temperature (500–1,100°C) a silicon compound reacts with hydrogen to form silicon on the material surface, followed by its inward diffusion.

Siliconizing by using physical vapor deposition (PVD) (evaporation, sputtering, etc.) of silicon onto a material followed by a suitable heat treatment under an inert gas was reported only recently (Vojtěch et al. 2008). This method produces thin, compact, and uniform Si-rich layers (see Fig. 2).

Liquid phase siliconizing (or silico-aluminizing) has been suggested for the surface protection of Ti-based intermetallics (Xiong et al. 2003). It consists of immersing a material into an appropriate melt, for example, Al-Si. Reaction of a substrate with a melt is relatively fast and produces thick layers composed of various Ti-Al-Si and Ti-Si phases.

Laser surface siliconizing involves rapid melting of a thin surface layer and a simultaneous feeding of powder silicon. As a result, a layer of the rapidly solidified (cooling rates of more than 10^4 K/s) Si-rich alloy with very fine microstructure is formed. The implantation of accelerated silicon ions onto the surface of a material can also serve as



Siliconizing, Fig. 2 Siliconized layer prepared by a combination of silicon electron beam evaporation and heat treatment at 900°C/2 h

a siliconizing method. However, applicability of both laser surface alloying and ion implantation is limited to research purposes, particularly due to their high cost. In addition, the reproducibility of these methods also appears problematic.

Mechanisms

When silicon diffuses from the surface to inner parts of an alloy, products of this process depend on substrate type, temperature, and time. An estimation of phase composition of a silicon-rich surface layer can be made using a particular equilibrium phase diagram. At the beginning of the diffusion process, silicon dissolves in a base metal to some extent. When its concentration in substrate achieves a certain critical value, the situation becomes favorable for the formation of a silicide in the substrate surface layer. Silicide containing the lowest concentration of Si precipitates first. On the opposite side, a base metal also diffuses to silicon on the surface and a Si-rich silicide may form when a maximum solubility is exceeded. The solid state reactions between silicon and substrate may thus convert them to a multi-layered structure composed of various silicides (see Fig. 3). The thickness of each sub-layer in the multi-layered structure is determined mainly by diffusion rates of a substrate metal and silicon through this layer. Generally, the faster the diffusion, the larger the thickness of a particular sub-layer.

Fe-Si, Ni-Si, and Ti-Si phase diagrams (Gale and Totemeier 2004) show that there are several types of binary silicides in each system. The situation becomes more complicated when an alloy of a complex chemical composition

is siliconized. In such a case, ternary or even more complex silicides may be formed. As stated above, silicides described in equilibrium phase diagrams do not appear in Si-rich surface layers in the same volume fractions. Some of them will be suppressed in their growth, while others will grow preferentially. It is influenced by diffusion rates and also by nucleation and growth parameters of silicides.

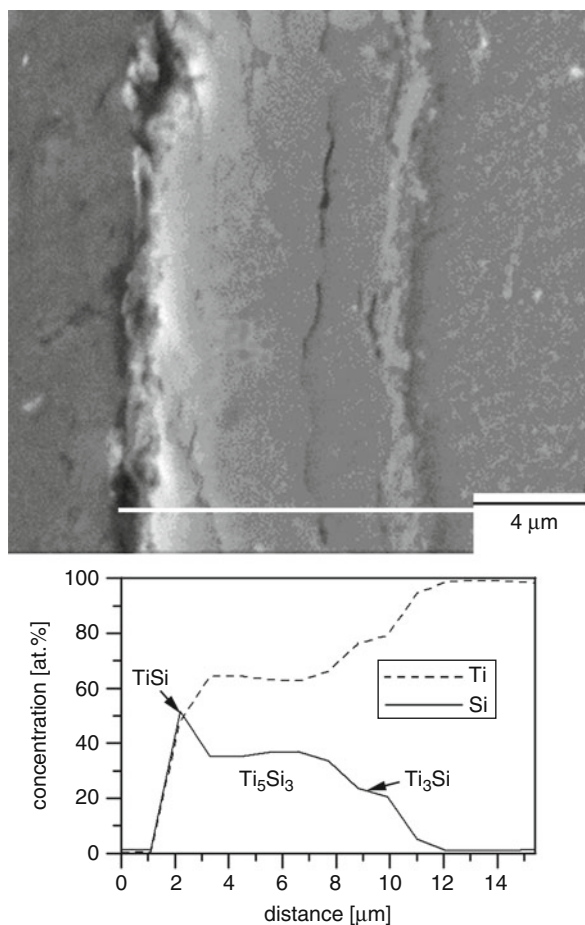
When low-carbon and low-alloy steels are siliconized, a surface layer may consist of Si-rich α -ferrite, Fe_3Si , and FeSi silicides (Cabrera and Kirner 1989). The general trend is that the increase in siliconizing temperature and time of treatment also increases the formation of higher (or Si-rich) silicides. A silicide case is hard, brittle, and corrosion resistant. Siliconizing of Cr-Ni austenitic stainless steels is more complicated. In contrast to nickel, which is a stabilizer of FCC austenite phase, silicon is a BCC ferrite stabilizer. This means that siliconizing results in transformation of the FCC austenitic matrix into a BCC structure in a surface layer. The siliconized case then consists of BCC or BCC + FCC matrix and iron silicides (Huang et al. 2006).

Nickel-base superalloys can also be successfully siliconized to improve their high-temperature performance. These alloys are characterized by complex chemical compositions. Beside nickel, they commonly contain chromium, titanium, iron, aluminum, molybdenum, and other additives. For this reason, phase composition of the siliconized case becomes relatively complicated. It may contain FCC NiCrSi matrix, binary nickel silicides (Ni_3Si , Ni_3Si_2 , NiSi), ternary, for example, nickel-chromium or nickel-titanium silicides, as well as other phases (Gründling and Bauer 1982).

Beside steels and Ni-based alloys, siliconizing of titanium alloys by the methods given before was previously already reported in a number of works. A siliconized case may consist of various titanium silicides (Ti_3Si , Ti_5Si_3 , Ti_5Si_4 , TiSi , TiSi_2), however, it is often observed that Ti_5Si_3 silicide with a high melting point above 2,000°C predominates (see Fig. 3).

Key Applications

Siliconizing of metals has been known since 1915. It is mainly applied to improve high-temperature oxidation resistance, hot corrosion resistance, carburization resistance, and wear resistance of steels, nickel-based alloys, titanium-based alloys, and other materials. Examples of the main applications include gas turbine blades, chemical reactor components, and furnace tubes in chemical and fossil fuel power plants. In these applications, metals are exposed to high temperatures and various oxidizing or



Siliconizing, Fig. 3 Multi-layered structure of a siliconized case on titanium prepared by pack cementation with elemental profiles along a *drawn line*. The case consists of at least three silicides but is dominated by Ti_5Si_3

non-oxidizing gaseous media containing oxygen, carbon compounds, sulfur compounds, and so on. In addition, more or less chemically aggressive fused salt deposits and erosive solid particles that mechanically damage the surface are in contact with metals. Siliconizing, combined with surface alloying methods like Ti-Si, Cr-Si, Ni-Cr-Si, Al-Si coatings, are applied to prolong the life of components.

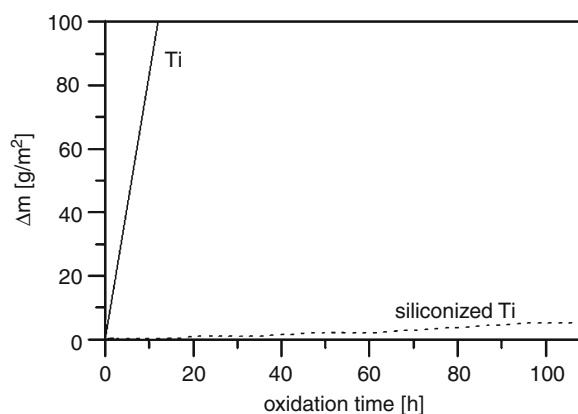
High-Temperature Oxidation Resistance

High-temperature oxidation occurs when a metal comes into contact with an atmosphere containing a sufficient amount of oxygen and/or other oxidizing compounds at a high temperature. During the early stage of oxidation, an

oxide scale forms due to a reaction of metal and oxygen. Further oxidation then requires inward transport of oxygen and/or outward transport of metal through the oxide scale. Oxidation rate is governed by the rate of this transport. When oxide scales are porous and poorly adherent, and when they spall off easily or are volatile, they do not represent an efficient barrier against further oxidation. In such a case, oxidation proceeds rapidly and material fails early in service. On the other hand, when scales are compact, adherent, and pore-free, transport of species through them becomes difficult and, as a result, oxidation rate progressively reduces with time. Such alloys are considered to be oxidation resistant.

Silicon is known to show a high affinity to oxygen. Therefore, the positive influence of siliconizing (i.e., the enrichment of a surface layer of a material with silicon) on high-temperature oxidation resistance is generally explained in terms of silicon dioxide (SiO_2) formation in scales during oxidation. Silicon dioxide may be either amorphous or crystalline, depending on temperature. It may be observed in scales as a separate sub-layer at a scale/metal interface or as a part of other oxide compounds. In all these cases, silicon dioxide represents a very effective barrier against inward penetration of oxygen. Even when it is dissolved in other oxides, it can retard the oxygen penetration towards the substrate considerably. Oxidation of titanium is a typical example. When titanium is protected by a layer of titanium silicide, oxidation may produce scales composed of a solid solution of SiO_2 in rutile (TiO_2). Small Si^{4+} cations incorporate into interstitial positions in a rutile lattice. To maintain electroneutrality, oxygen vacancy concentration reduces. Oxygen vacancies are responsible for oxygen diffusion through rutile and, therefore, diffusion rate decreases. In nickel alloys, formation of Ni_2SiO_4 in scales during oxidation is very effective in limiting transport of species.

Carbon steels and low-alloy steels generally show a poor oxidation resistance at high temperatures. Oxide scales are porous and are not an effective barrier against further oxidation. It was shown in a number of reports that siliconizing significantly improves high-temperature oxidation resistance of these materials. It was, for example, reported that siliconizing of 1010 carbon steel by chemical vapor deposition in a mixture of SiH_4 and H_2 is very effective (Cabrera and Kirner 1989). The steel was siliconized to achieve a weight gain of $0.16 \text{ mg Si}\cdot\text{cm}^{-2}$. This value corresponded to a Fe_3Si silicide case of about $1 \mu\text{m}$ in thickness. This case provided a reduction of oxidation rate at 800°C measured as weight gain in $\mu\text{g}\cdot\text{cm}^{-2} \text{ min}^{-1}$ even by a factor of $2\cdot 10^4$. Unfortunately, carbon steels are



Siliconizing, Fig. 4 Oxidation kinetics of pure titanium and titanium siliconized by pack cementation at 1,100°C/3 h (oxidation at 850°C in air) expressed as a specific weight gain Δm versus oxidation time (Vojtěch et al. 2006)

not typically used at such high temperatures because they have low high-temperature strength.

Sufficient creep resistance and high-temperature strength are achieved for high-alloyed steels and, particularly, for Ni-based superalloys. These materials commonly contain high concentrations of chromium, which is very beneficial for high-temperature oxidation resistance because it improves protective effect of scales. Siliconizing leads to additional improvement of oxidation resistance of these materials at high temperatures. Since oxidation resistance of untreated high-alloy steels and Ni-based alloys is much better than that of untreated carbon steels, the effect of siliconizing on these materials is smaller than on carbon steels. When 430 stainless steel containing about 17%Cr was siliconized by CVD to a value of 0.5 mg Si·cm⁻², oxidation rate at 1,000°C reduced by a factor of 7 (Cabrera and Kirner 1989). Another example was a 310 stainless steel (26% Cr, 19% Ni) treated by pack cementation. Specific weight gains after oxidation at 800°C/48 h for untreated and siliconized steel were 4.19 and 2.96 mg·cm⁻², respectively (Hsu and Tsai 2000).

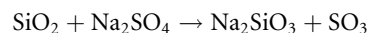
Titanium alloys also oxidize rapidly at above 600°C. TiO₂ scales are porous, stratified, and suffer from spallation during thermal cycling. Siliconizing by pack cementation at 1,100°C/3 h is able to reduce oxidation rate at 850 °C in air considerably, as is illustrated in Fig. 4. Weight gain of pure titanium after 10 h of oxidation reached about 80 g m⁻², whereas 100-h oxidation of siliconized titanium resulted in a low weight gain of about 5 g m⁻². A similar effect was observed for titanium protected by a thin (1–2 μm) silicide case prepared by a combination of

physical vapor deposition and heat treatment. Here, oxidation kinetics at 900°C was measured as a scale thickness versus oxidation time. Scales on pure titanium grew very rapidly and reached a thickness of 300 μm after 30 h of oxidation. In contrast, 100-h oxidation of the siliconized titanium produced scales with a thickness of only 20 μm (Vojtěch et al. 2008). Similarly to carbon steels, common titanium-based alloys are not applied at high temperatures above 550 °C because of their poor creep resistance and low strength. Fortunately, Ti-based intermetallics (TiAl, Ti₃Al) are characterized by much higher operation temperatures, exceeding 700°C. Siliconizing of these materials may significantly prolong their life in oxidizing environments (Liang and Zhao 2001).

Hot Corrosion Resistance

In high-temperature applications, metals are often exposed to gas-molten salt environments. Siliconizing may result in an improvement of corrosion resistance in these media. Similarly to the high-temperature oxidation, this improvement is generally attributed to the silicon dioxide (SiO₂)- and/or silicate (e.g., Ni₂SiO₄)-containing scales. However, corrosion behavior of silicon dioxide in gas-molten salt environments depends on the chemical nature of molten salt deposits and on oxygen partial pressure. Generally, formation of silicon dioxide is supported by a sufficient amount of oxygen in surrounding atmosphere. When oxygen partial pressure falls below a critical value, volatile SiO may form instead of solid SiO₂, which has no protective effect and material is corroded rapidly.

Deposits in industrial gas turbines often contain condensed alkaline sulfates (Na₂SO₄), which result from reactions of sulfur compounds, oxygen, and alkalis. Alkaline sulfates may react with silicon dioxide to form silicates, accompanied by a decrease in melting point of protective scales and by a reduction of its protective effect:



When thin layers of sulfates are present and oxygen and sulfur oxide partial pressures are sufficient, the reaction above is slow. On the other hand, when deposits are thick to suppress diffusion of oxygen to the melt-SiO₂ interface, gaseous SiO may form, which reacts with sulfate to produce silicates and accelerates corrosion.

Stability of SiO₂ also depends on acidity or basicity of salt melts. SiO₂ scales are usually inert to acid and neutral salt melts, whereas their corrosion is accelerated by presence of basic salts in which silicon dioxide readily dissolves.

Carburization Resistance

Carburization process may occur in industrial chemical syntheses in which metals (usually steels) come into contact with hydrocarbons at high temperatures. In these processes carbon is formed as a byproduct that may diffuse into material and negatively influence its properties. Siliconizing has been shown to protect against carburization. When the amount of silicon in a surface layer is sufficient, silicon dioxide forms in scales in oxidizing environments. SiO_2 is an excellent barrier against carburization. In addition, silicon present in solid solution in a base metal also reduces carburization rate because it decreases carbon solubility and diffusivity in Fe-based matrix (Southwell et al. 1987).

Wear Resistance

Siliconizing generally improves resistance of metallic materials to wear. It is attributed to a hardening of metal due to the presence of silicon in a surface layer. When silicon is present in a solid solution in a base metal, like in siliconized steels (see above), solid solution hardening occurs due to a stress in crystal lattice induced by silicon atoms. In such a stressed lattice dislocation motion is hindered. When silicon forms silicides in a surface layer, it leads to further significant hardening. Silicides are characterized by strong chemical bonds and often ordered structures in which plastic deformation is not possible. For these reasons, they are very hard and brittle. As an example, hardness of FeSi, NiSi, TiSi, and Ti_5Si_3 silicides are about 900, 700, 1,000, and 1,500 HV, respectively.

Improvement of wear resistance by siliconizing is the most significant for soft materials (i.e., for low-carbon steels). Siliconized case on low-carbon steels generally achieves hardness between about 300 and 600 HV. For titanium, surface hardness as high as 1,500 HV may be achieved.

References

- R. Bianco, M.A. Harper, R. Rapp, Codepositing elements by halide activated pack cementation. *JOM* **43**, 20–25 (1991)
- A.L. Cabrera, J.F. Kirner, Formation of silicon diffusion coatings on ferrous alloys from their reaction with silane. *Surf. Coat. Technol.* **39/40**, 43–51 (1989)
- W.F. Gale, T.C. Totemeier, *Smithells Metals Reference Book*, 8th edn. (Elsevier, Amsterdam, 2004)
- H.W. Gründling, R. Bauer, The role of silicon in corrosion-resistant high temperature coatings. *Thin Solid Films* **95**, 3–20 (1982)
- H.W. Hsu, W.T. Tsai, High temperature corrosion behaviour of siliconized 310 stainless steel. *Mater. Chem. Phys.* **64**, 147–155 (2000)
- H.L. Huang, T.Y. Lee, D. Gan, The microstructure of siliconized type 310 stainless steel. *Mater. Sci. Eng.* **A422**, 259–265 (2006)
- C. Klam, J.P. Millet, H. Mazille, J.M. Gras, Chemical vapor deposition of silicon onto iron and steel substrates: oxidation and corrosion properties of coated materials. *Mater. Manuf. Process.* **6**, 451–467 (1991)

- W. Liang, X.G. Zhao, Improving the oxidation resistance of TiAl based alloy by siliconizing. *Scripta Mater.* **44**, 1049–1054 (2001)
- G. Southwell, S. Macalpine, D.J. Young, Silicide coatings for carburization protection. *Mater. Sci. Eng.* **88**, 81–87 (1987)
- D. Vojtěch, T. Kubatík, M. Pavlíčková, J. Maixner, Intermetallic protective coatings on titanium. *Intermetallics* **14**, 1181–1186 (2006)
- D. Vojtěch, P. Novák, P. Macháček, M. Morťaniková, K. Jurek, Surface protection of titanium by Ti_5Si_3 silicide layer prepared by combination of vapor phase siliconizing and heat treatment. *J. Alloy. Comp.* **464**, 179–184 (2008)
- H.P. Xiong, Y.H. Xie, W. Mao, Improvement in the oxidation resistance of the TiAl-based alloy by liquid-phase siliconizing. *Scripta Mater.* **49**, 1117–1122 (2003)

Simplified EHL Solution Methods

G. E. MORALES-ESPEJEL

SKF Engineering & Research Centre, Nieuwegein,
The Netherlands

Université de Lyon, CNRS INSA–Lyon, LaMCoS
UMR5259, Villeurbanne, France

Synonyms

Analytical EHL solution methods; Semi-analytical EHL solution methods

Definition

Simplified elastohydrodynamic lubrication (EHL) solution methods are calculation methods in EHL for film thickness and pressures, which make use of simplifications in either the Reynolds equation or contact geometry to achieve a semi-analytical approximated solution, resulting in much faster computing times compared with full numerical solutions. Such methods exist for both smooth- and rough-surface EHL problems.

Scientific Fundamentals

In the past, numerical solutions of the EHL problem represented a challenge for engineers due to convergence demands and high cost of computers. They only became practical a few decades ago. Therefore, engineers developed alternative simplifications to the problem in order to produce faster solutions. Examples for smooth surfaces are the works of Ertel (1949), also published under the name of Grubin (Grubin et al. 1949), followed by many other authors, e.g., Greenwood (1972), (Greenwood and Morales-Espejel 1995) and Christensen (1979), where simplifications are applied mainly to the geometry of the EHL contact, and the inlet, center, and outlet regions are solved separately. This methodology allows for the

consideration of the proper assumptions and simplifications in each region of the contact. Simplified methods for rough surfaces in EHL (micro-EHL) are based on the linearization of the Reynolds equation and the use of solutions for sinusoidal surfaces. Works along this line include Greenwood and Johnson (1992), Greenwood and Morales-Espejel (1994), Venner and Lubrecht (1999), and Hooke (1998). The idea is that once the fundamental solution for deformation and hydrodynamic pressures is obtained, this can be generalized to “any” real roughness with the use of the discrete Fourier transform (DFT).

Detailed descriptions of some of the methods are given next.

Smooth Surface Methods

The Ertel–Grubin model considers the following assumptions: the contact is infinitely long in the transverse direction (line contact), the lubricant follows Newtonian behavior and the Barus law, the fluid is incompressible, the Reynolds equation applies, and the contact dimensions are small in comparison with those of the contacting bodies, thus half-space elasticity applies. Under these conditions, the Reynolds equation is

$$\frac{dp}{dx} = \frac{12\bar{u}\eta}{h^3}(h - h^*) \quad (1)$$

where p = pressures, h = film thickness at location x , h^* = film thickness value at the point where $dp/dx = 0$, \bar{u} = mean lubricant entrainment speed, x = coordinate along the rolling direction, and η = lubricant viscosity under the contact at location x . Next, the following dimensionless groups are introduced:

$$X = x/a, \hat{H} = 2R_x h/a^2, P = p/p_o, \\ Q = q/p_o, K = 48\bar{u}\eta_o R_x^2/(a^3 p_o)$$

with p_o = maximum Hertzian pressure, a = Hertzian contact semi-width along x , R_x = radius of curvature in x direction, q = reduced pressure, $q = [1 - \exp(-\alpha p)]/\alpha$, α = lubricant viscosity-pressure coefficient in the Barus law ($\eta = \eta_o \exp[\alpha p]$), and η_o = lubricant viscosity at ambient conditions.

Thus, the dimensionless Reynolds equation becomes

$$\frac{dQ}{dX} = K \frac{\hat{H} - \hat{H}^*}{\hat{H}^3} \quad (2)$$

where dimensionless pressure P can be calculated from the reduced dimensionless pressure Q ,

$$P = -\frac{1}{K_1} \ln(1 - K_1 Q) \quad (3)$$

where $K_1 = \alpha p_o$.

For the calculation of Q one needs to integrate (2),

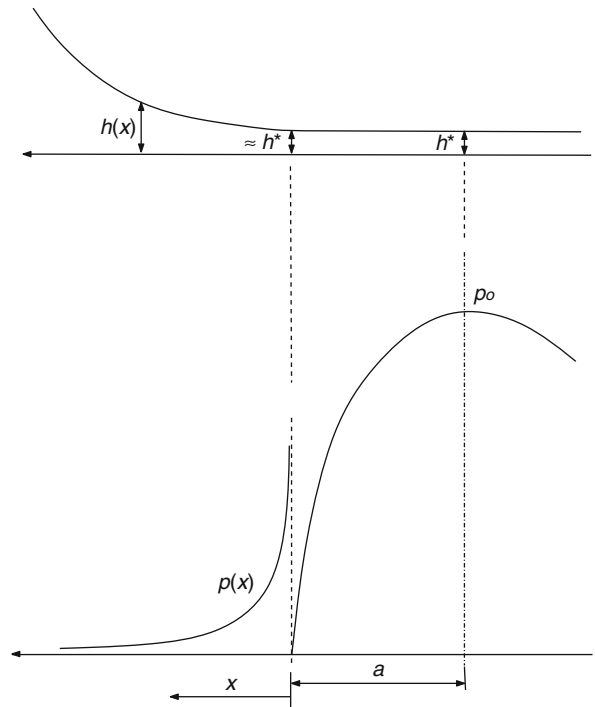
$$Q = Q_s + K \int_{X=1}^X \frac{\hat{H} - \hat{H}^*}{\hat{H}^3} dX \quad (4)$$

Q_s is the value of Q at $X = 1$. The correct axis in the contact is X negative to the left of the contact center (inlet) and positive to the right; however, for simplicity of signs in the integrations at the inlet $-X$ will be replaced by X in the whole of this section, see Fig. 1. The calculation of the pressures Q and the film thickness \hat{H}^* (originally at location $dP/dX = 0$, but under the assumption of parallel film, now at location $X = 1$ where $\hat{H}^* = \hat{H}_c$) can be done by imposing boundary conditions.

For the calculation of the reduced pressures, Ertel–Grubin introduced the following assumption, $P \rightarrow \infty$ at $X = 1$, which becomes $q_s = 1/\alpha$ or $Q_s = 1/K_1$ at $X = 1$. This is a good approximation for any large P (not necessarily $P = \infty$).

Solution for \hat{H}^*

In the calculation of \hat{H}^* , Ertel–Grubin introduced the following condition, $P = 0$ for $X = \infty$ (in a numerical scheme, this can be approximated to large values of X , namely $X = L$, L being a sufficiently large number to



Simplified EHL Solution Methods, Fig. 1 Inlet geometry for the Ertel–Grubin model in dimensional quantities

avoid “numerical starvation”). Substituting this in (4) and $Q_s = 1/K_1$ at $X = 1$ gives

$$0 = \frac{1}{K_1} - K \int_{X=1}^{\infty} \frac{\hat{H} - \hat{H}^*}{\hat{H}^3} dX \quad (5)$$

At this stage, (5) can be evaluated numerically for \hat{H}^* (as originally done by Ertel–Grubin, see also (Morales-Espejel et al. 2003)) using the Hertz shape for the inlet.

$$\hat{H} - \hat{H}^* = |X|(X^2 - 1)^2 - \ln[|X| + (X^2 - 1)^{1/2}] \quad (6)$$

However, a fully analytical solution for \hat{H}^* and P is also possible by following the scheme of Wolveridge et al. (1970–71) (originally developed by Dowson and Higginson), with the use of Crook’s remarkably accurate approximation for the inlet region close to the Hertzian zone,

$$\hat{H}(X) = \hat{H}^* + C(X - 1)^{3/2} \quad (7)$$

$$\text{with } C = \frac{4\sqrt{2}}{3}.$$

Substituting (7) into (5), yields

$$0 = \frac{1}{K_1} - K \int_{X=1}^{\infty} \frac{C(X - 1)^{3/2}}{[\hat{H}^* + C(X - 1)^{3/2}]^3} dX \quad (8)$$

by taking $\sigma = (X - 1)(C/\hat{H}^*)^{2/3}$, the integral in (8) can be written for any point X in the upper limit as:

$$I(X) = \frac{1}{(\hat{H}^*)^{4/3} C^{2/3}} \int_{\sigma=0}^{[(C/\hat{H}^*)^{2/3}(X-1)]} \frac{\sigma^{3/2}}{(1 + \sigma^{3/2})^3} d\sigma \quad (9)$$

which can be solved with the following approximation:

$$\int_0^{\infty} \frac{\sigma^{3/2}}{(1 + \sigma^{3/2})^3} d\sigma = 0.2687$$

Therefore, substituting this in (9) and solving for \hat{H}^* one obtains:

$$\hat{H}^* = \frac{(K_1 K \hat{B})^{3/4}}{C^{1/2}} \quad (10)$$

finally with $C = 1.88$ and $\hat{B} = 0.2687$

$$H^* = 0.2722(K_1 K)^{3/4} \quad (11)$$

Solution for the Inlet Pressures

Once \hat{H}^* is known, it can be substituted into (4) and the reduced pressures $Q(X)$ can be calculated numerically. Alternatively, analytically solving for K (5), substituting it into (4), and transforming back Q into P with the use of (3) leads to

$$P(X) = -\frac{1}{K_1} \ln[I(X)/I(L)] \quad (12)$$

with $I(L)$ and $I(X)$ given both by (9) with different upper limits.

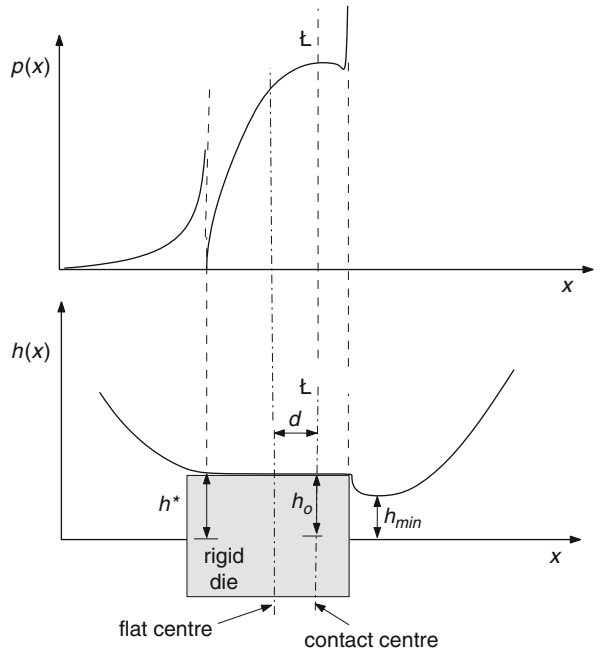
Off-Center Flat Solution

Ertel–Grubin made an attempt to calculate the location and the values of the outlet EHL pressure spike by using a combination of the pressures produced by a rigid die contact and the Hertzian solution. However, this did not produce minimum film thickness values or the exact location of the pressure spike.

Greenwood (1972) introduced an extension of the Ertel–Grubin method to calculate the left-hand side of the outlet pressure spike and the height of the outlet constriction, thus finally allowing for the calculation of the minimum film thickness in EHL contacts. The method is based on the assumption that the shape and pressures in this region can be approximated using the elastic solution of an off-center Hertzian flat, equivalent to an off-center rigid die, see Fig. 2.

The elastic pressure distribution of the central region is then given by

$$\frac{p}{p_o} = \frac{1}{(1 + 2d/a)^{1/2}} \left[(1 - \xi^2)^{1/2} + (d/a) \left(\frac{1 + \xi}{1 - \xi} \right)^{1/2} \right] \quad (13)$$



Simplified EHL Solution Methods, Fig. 2 Off-center geometry for the outlet Greenwood’s extension of the Ertel–Grubin model

where d = off-center distance, $\xi = x/b$. The resulting profile after elastic deformation is

$$\begin{aligned} h &= [a^2/(4R_x)]A(\xi) & \text{for } \xi < -1 \\ h &= 0 & \text{for } -1 < \xi < 1 \\ h &= [a^2/(4R_x)]B(\xi) & \text{for } \xi > 1 \end{aligned} \quad (14)$$

with,

$$\begin{aligned} A(\xi) &= 2 \left[|\xi|(\xi^2 - 1)^{1/2} - \cosh^{-1}|\xi| + 2(d/a) \{ \xi^2 - 1 \}^{1/2} - \cosh^{-1}|\xi| \right] \\ B(\xi) &= 2 \left[\xi(\xi^2 - 1)^{1/2} - \cosh^{-1}\xi - 2(d/a) \{ (\xi^2 - 1)^{1/2} + \cosh^{-1}\xi \} \right] \end{aligned}$$

The Ertel–Grubin analysis for the inlet of the contact can be repeated here using the inlet shape given by (14). This will produce the following equation:

$$\frac{a^2}{8\alpha\eta_o\bar{u}R_x^2} = 24 \int_0^\infty \frac{A(\theta) \sinh \theta d\theta}{[H_o + A(\theta)]^3} \quad (15)$$

where $A(\theta) = \sinh(2\theta) - 2\theta + 4(d/a)(\sinh \theta - \theta)$, $\xi = -\cosh \theta$ and $H_o = 4R_x h^*/a^2$. This integral can be evaluated numerically for given values of H_o and d/a .

Consider now the outlet, applying the Reynolds boundary condition (i.e., $q = 0$ when $h = h^*$) after the constriction, therefore

$$\frac{a^2}{8\alpha\eta_o\bar{u}R_x^2} = 24 \int_0^{\theta^*} \frac{B(\theta) \sinh \theta d\theta}{[H_o + B(\theta)]^3} \quad (16)$$

with $B(\theta) = 4(d/a)(\sinh \theta + \theta) - [\sinh(2\theta) - 2\theta]$ and θ^* is the value at which the film thickness returns to h^* , where $B(\theta^*) = 0$. Again, numerical evaluation is possible for given values of H_o and d/a . Finally, the solution of the problem implies the finding of these two constants for given operating conditions. This is achieved by iterations until the two integrals (15) and (16) are equated. Figure 3 shows some solutions obtained by reference (Greenwood 1972).

The major limitation of this approach is that the outlet constriction is calculated without the consideration of the outlet hydrodynamic pressures (right-hand side of the pressure spike). Therefore, this might result in slightly lower minimum film thicknesses.

Outlet Pressures and Complete Solution

Greenwood and Morales-Espejel (1995), (Morales-Espejel and Wemkamp 2008) used the concept of linear fracture mechanics to develop an approach for the elastic displacement and pressure calculation in EHL. This concept combined with the use of the Reynolds equation in a similar fashion as used by Ertel–Grubin allows for the calculation of the inlet, central region, and outlet geometries and pressures in an EHL contact. Therefore, the outlet hydrodynamic pressures can be calculated in addition to its

effects on the outlet shape. Complete line-contact EHL solutions can be reconstructed with this scheme.

The only requirement for the application of elastic fracture mechanics equations is that in the cracked body the displacements on the line beyond the crack tip should be zero. By considering a dry Hertzian contact it is possible to see that, once the Hertzian deformation has occurred, any external traction applied outside the contact must produce zero displacements along the contact line, however, it will produce internal stresses $\sigma(x)$ similarly as a crack. Therefore, either a Hertzian or an EHL contact, in this sense, have the same geometry as a doubly cracked body, see Fig. 4.

The final internal stresses $\sigma(x)$ in the body represent the final pressures in the EHL central zone of the contact, p_R represents the outlet EHL pressures and p_L the inlet EHL pressures. After (Greenwood and Morales-Espejel 1995), the final internal stresses are given by

$$\begin{aligned} \sigma(x) \approx & -Q\sqrt{a^2 - x^2} - M_L\sqrt{\frac{a-x}{a+x}} - \Delta M\sqrt{\frac{a+x}{a-x}} \\ & - \frac{\sqrt{a^2 - x^2}}{\pi} \int_{x_1=a}^{L_R} \frac{p_R(x_1)dx_1}{(x_1-x)\sqrt{x^2 - a^2}} \end{aligned} \quad (17)$$

and the outlet elastic displacements $v(x)$ can be calculated from the outlet slopes for given $p_R(x)$ and $p_L(x)$.

$$\begin{aligned} \frac{E'}{2} \left(\frac{\partial v}{\partial x} \right)^R &= -\Delta M \sqrt{\frac{x+a}{x-a}} + \frac{1}{\pi} \sqrt{x^2 - a^2} \\ & \left[\int_{x_1=a}^{L_R} \frac{p_R(x_1)dx_1}{(x_1-x)\sqrt{x_1^2 - a^2}} - \int_{x_2=a}^{L_L} \frac{p_L(x_2)dx_2}{(x_2+x)\sqrt{x_2^2 - a^2}} \right] \end{aligned} \quad (18)$$

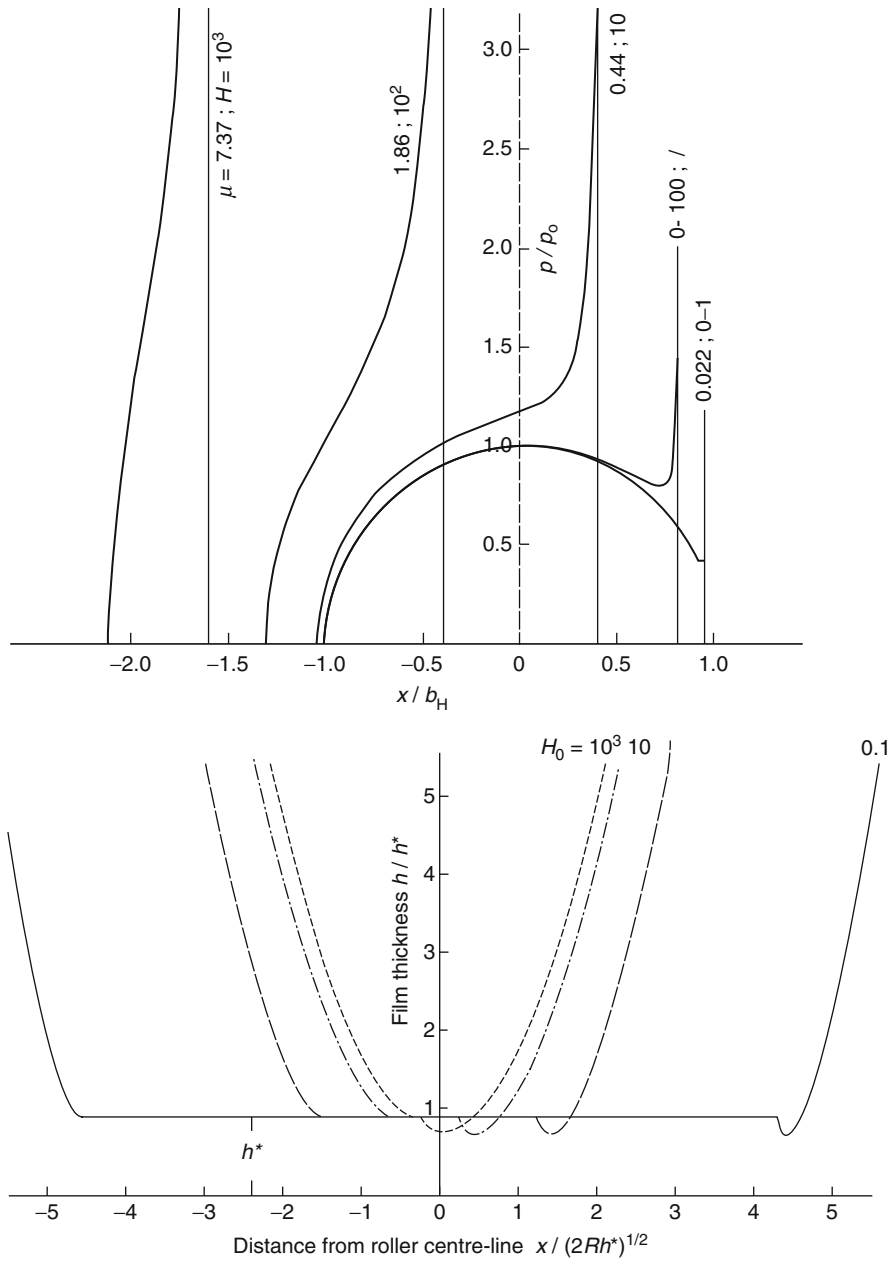
where $Q = E'/(4R_x)$, M' s are stress concentration factors, defined as

$$\begin{aligned} M_L &= \frac{1}{\pi} \int_{-a}^{L_L} \frac{p_L(x)}{\sqrt{x^2 - a^2}} dx; \quad M_R = \frac{1}{\pi} \int_a^{L_R} \frac{p_R(x)}{\sqrt{x^2 - a^2}} dx; \\ M_h &= Q(b-a); \quad \Delta M = M_h - M_R \end{aligned}$$

with a being the semi-width of the flat zone in the EHL contact and b the semi-width of an original Hertzian contact problem where the flat is included, to be found in the overall equilibrium of the pressures. L_L = large distance for which $p_L \approx 0$, and L_R = large distance for which $p_R(x) \approx 0$.

Once the slope is calculated, it can be integrated to obtain the elastic displacements due to the inlet and outlet pressures. The final displacements are then obtained by adding the Hertzian displacements,

$$v_f(x) = v(x) + v_h(x) \quad (19)$$



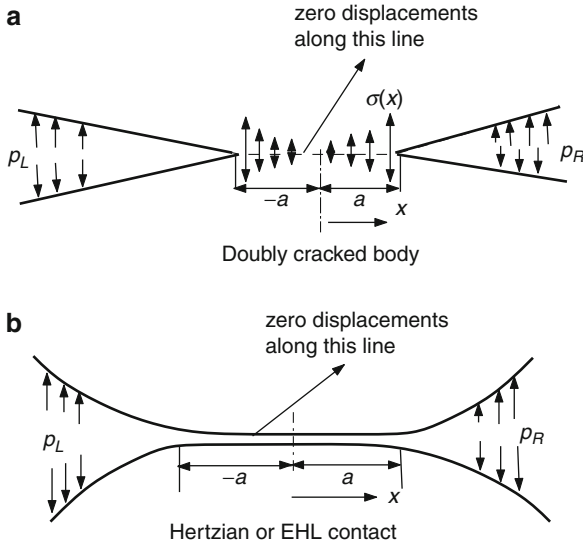
Simplified EHL Solution Methods, Fig. 3 Pressure distribution and elastic profile for an off-center flat solution in EHL as given by Greenwood (1972), in the reference's nomenclature $b_H = a$

$$v_h(x) = \frac{a^2}{2R_x} \left\{ \frac{x}{a} \left[\left(\frac{x}{a} \right)^2 - 1 \right]^{1/2} - \cosh^{-1} \left(\frac{x}{a} \right) \right\} \quad (20)$$

Finally, the computational procedure for full EHL solutions can be summarized as follows:

1. Use dimensionless parameters $X = x/a$, $V(X) = 2R_x v/(a^2)$ and $P = \alpha p$ to obtain

$$K_1 = \frac{\alpha E' a}{4R_x}; \quad C = \frac{a^3}{48\alpha\eta_o \bar{u} R_x^2}; \quad H^* = \frac{4Rh^*}{b_h^2}$$



Simplified EHL Solution Methods, Fig. 4 Geometrical similarities between a doubly cracked body and a Hertzian or EHL contact

b_h being the semi-width of the equivalent Hertzian contact for the same total load.

2. Choose operating conditions K_1 and H^* . Finally, Guess $\alpha\Delta M$, Assuming $P_L(X) = 0$ to Obtain C .
3. Solve inlet, including effect of $P_R(X)$ to obtain C_i . Adjust $\alpha\Delta M$ to make $C_i = C$.
4. Find a new outlet solution allowing for the effects of the inlet pressure.
5. Repeat steps 2 and 3 until convergence is reached, then find internal stresses and total load.

Greenwood and Morales-Espejel (1995) includes some full examples reproduced here in Fig. 5, with the operating conditions given as Moes parameters, M and L ,

$$M = \frac{w}{E'R_x} \left(\frac{2E'R_x}{\eta_0 \bar{u}} \right)^{1/2}; \quad L = \alpha E' \left(\frac{2E'R_x}{\eta_0 \bar{u}} \right)^{-1/4}$$

Rough-Surface Methods

Greenwood and Johnson (1992) first noticed that, in EHL, sinusoidal roughness would produce nearly sinusoidal pressure ripples, and in the presence of sliding using a Newtonian fluid, the roughness is largely flattened and replaced by large pressure ripples. They worked out simple relationships to describe pressure and deformed roughness. Greenwood and Morales-Espejel (1994) developed a methodology to generalize this concept to any roughness by linearizing the Reynolds equation. This methodology is based on the observations of Venner (1991) that at the

center of a low-amplitude wavy EHL contact, the lubricant viscosity is so high that the Poiseuille flow term in the Reynolds equation can be disregarded. Reducing it to a simple linear transport equation,

$$\bar{u} \frac{\partial(\rho h)}{\partial x} + \frac{\partial(\rho h)}{\partial t} = 0 \quad (21)$$

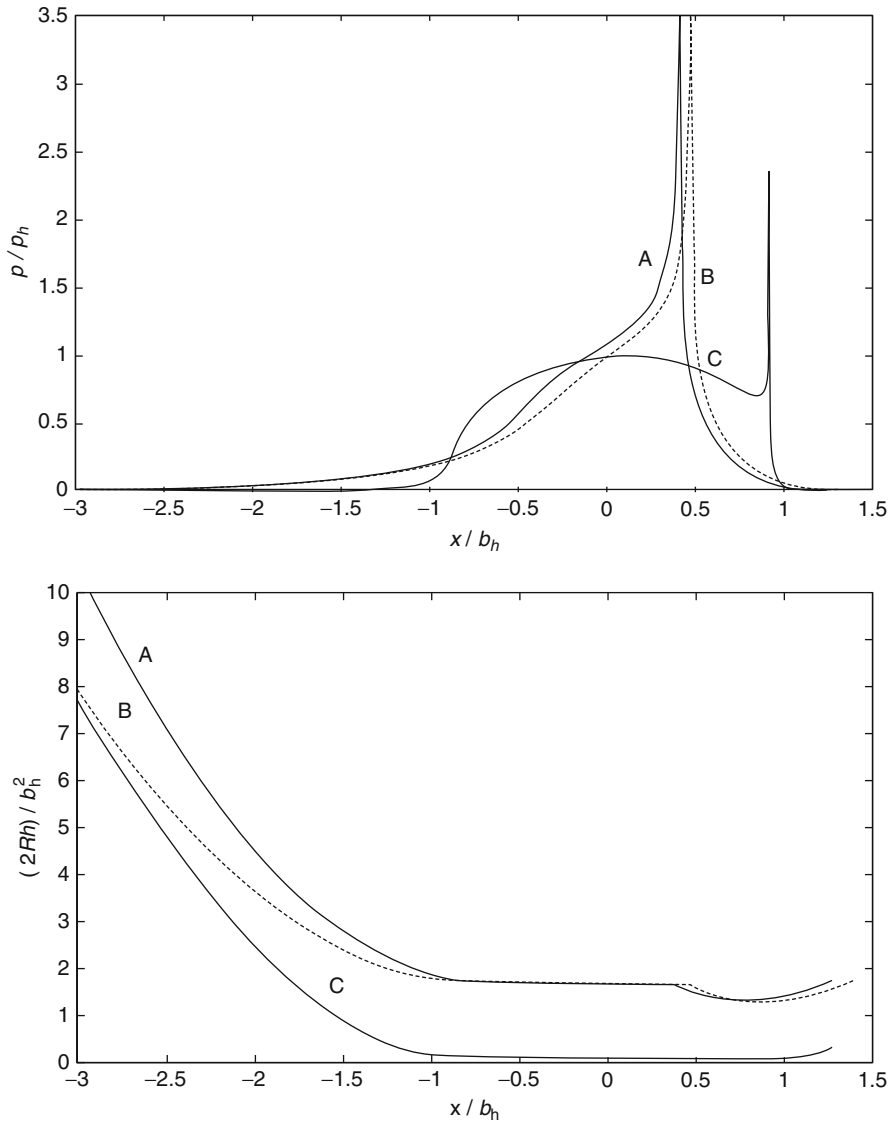
Newtonian Fluid

The solution for micro-EHL pressure and local film fluctuations with transverse roughness in a line contact and with Newtonian fluid is described in detail in reference (Greenwood and Morales-Espejel 1994). Where the complete transient solution of the problem is shown to be made of two components: (1) the particular integral, i.e., steady state solution of the problem, and (2) a complementary function, i.e., flow excitation at the inlet of the contact produced by the incoming (partly deformed) roughness and solution of the homogeneous equation (21), see Fig. 6, which shows a case of rolling/sliding. The reference shows equations for the two solutions considering an initial sinusoidal roughness. However, the amplitude of the complementary function cannot be obtained with the original method of reference (Greenwood and Morales-Espejel 1994) because the inlet is not included in the formulation. Morales-Espejel et al. (2003) used the “amplitude reduction” numerical results from Venner and Lubrecht (1999) to calculate this amplitude. Because in Newtonian conditions the complementary function dominates the deformed clearance, this deformation is controlled by a new variable named ∇ .

$$\nabla = (\lambda/a) M^{0.5} L^{-0.5} = \sqrt{\frac{\pi}{3}} \frac{\lambda p_0^{1/2}}{\sqrt{\eta_0 \bar{u} \alpha R_x^{1/8} E'^{1/4}}} \quad (22)$$

For three-dimensional (3D) contacts and 3D micro-geometry, it is possible to extend the scheme by considering the central part of the contact only (Morales-Espejel et al. 2003), thus exploiting the similarities in the transverse and longitudinal amplitude reduction curves of Venner and Lubrecht (1999). The method applies a Fourier decomposition of the 3D topography, resolves for local film thickness and pressure fluctuations of each component, and then reassembles the results. A complete summary of the methodology is given in Gabelli et al. (2008).

Knowing that the pressure perturbations particularly depend on the relative compressibility of the lubricant and the flexibility of the surface material, the following equation is obtained for the prediction of



Simplified EHL Solution Methods, Fig. 5 Pressure distribution and elastically deformed profile for a complete semi-analytical solution in EHL as given by Greenwood and Morales-Espejel (1994), with b_h = Hertzian semi-width, $p_h = p_o$. [A] – $M = 1057$, $L = 8.45$; [B] – $M = 1189.5$, $L = 8.45$; [C] – $M = 2469$, $L = 10.06$

pressure ripple amplitude due to transverse sinusoidal waviness:

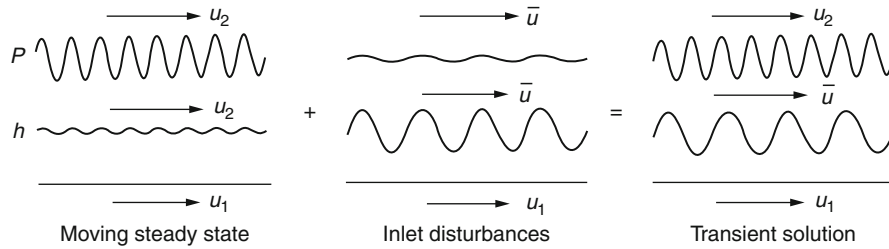
$$M^{-0.5} L^{0.5} E_c^{-1} \frac{\Delta P_x}{Z_1} = \frac{\pi}{C_0 + \nabla_x} \left(\frac{\Delta H_x}{Z_1} - 1 \right) \left(\cos \left[\frac{2\pi}{\lambda \bar{u}/(u_2 a)} (X - T) + \theta \right] - \cos \left[\frac{2\pi}{\lambda/a} (X - T) \right] \right) \quad (23)$$

and for the longitudinal sinusoidal waviness,

$$M^{-0.5} L^{0.5} E_c^{-1} \frac{\Delta P_y}{Z_1} = \frac{\pi}{C_0 + \nabla_y} \left(\frac{\Delta H_y}{Z_1} - 1 \right) \cos \left[\frac{2\pi}{\lambda/a} (X - T) \right] \quad (24)$$

∇_x (evaluated using ω_x) for transverse waviness and ∇_y (evaluated using ω_y) for longitudinal waviness, with $\Delta H_x/Z_1$ and $\Delta H_y/Z_1$ given by amplitude reduction curves of Venner and Lubrecht. The compressibility being

$$C_0 = \frac{\pi E'}{2a} M^{0.5} L^{-0.5} (\gamma_1 - \beta_1) \bar{h}$$



Simplified EHL Solution Methods, Fig. 6 Micro-EHL kinematics for a sinusoidal waviness as proposed by reference (Greenwood and Morales-Espejel 1994) for general rolling/sliding with Newtonian fluid

As it can be seen from (23) and (24) that the variable ∇ , initially proposed for line contacts and later extended for point contacts (Venner et al. 1999), is the key parameter in determining deformation, local film, and pressure fluctuations in micro-EHL. Using perturbation methods, Hooke (1998) has developed a fully analytical solution for sinusoidal waviness (extended to real roughness) that does not need the numerical results from Venner and Lubrecht. This analytical solution predicts “amplitude reduction” curves similar to those that are numerically obtained.

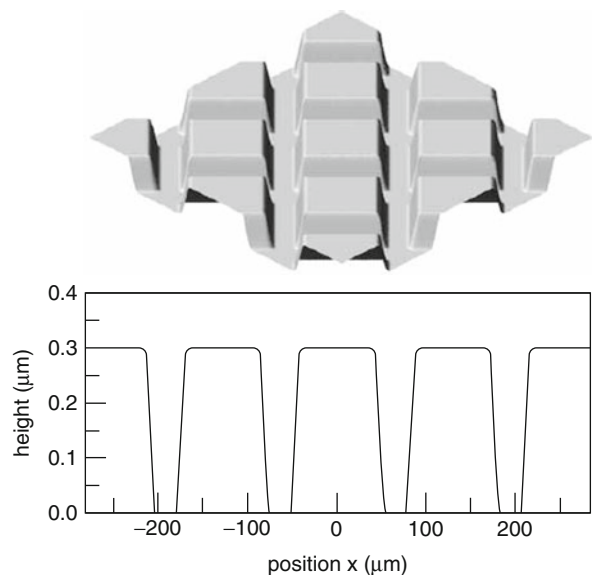
Non-Newtonian Fluid

When the rheology of a non-Newtonian fluid is considered in the presence of sliding, several aspects of the physics change. The fluid viscosity needs to account for shear thinning effects. The contribution of the particular integral in the local film is not negligible anymore; a considerable amplitude will remain. Thus, the amplitude reduction curves cannot directly be used to account for the missing amplitude in the complementary function. The complementary function will decay in amplitude as it enters into the contact and will propagate at the average speed of the surfaces.

Hooke et al. (2007) have developed a rapid method for the calculation of clearance and pressure fluctuations for 3D topography based on ideas similar to those described for Newtonian fluid. However, taking into account the non-Newtonian behavior of the lubricant, the amplitude of the complementary function is calculated by interpolation from perturbation method solutions. The rheology of the fluid is an Eyring relationship

$$\dot{\gamma} = \frac{\tau_0}{\eta} \sinh \left(\frac{\tau}{\tau_0} \right) \quad (25)$$

Hooke et al. (2007) gives some examples of rolling/sliding solutions for “deterministic” real surfaces as produced in experiments; Fig. 7 shows an example. The solutions for pressure and deformed shape for different sliding/rolling ratios are given in Fig. 8, v being the

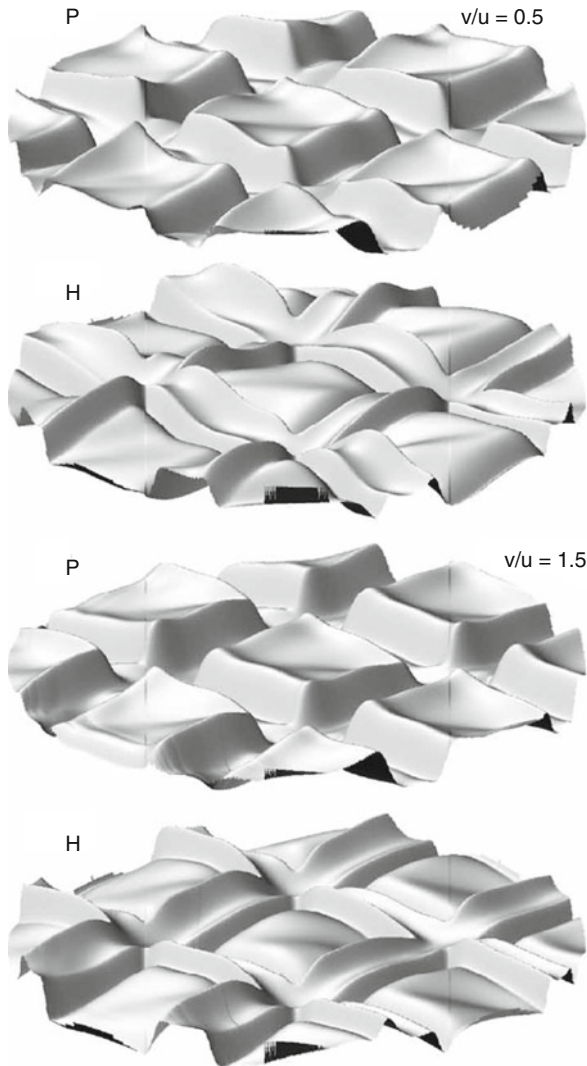


Simplified EHL Solution Methods, Fig. 7 Asperity geometry used as input in the example given by Hooke et al. (2007)

sliding velocity and u the entrainment speed of the lubricant in the contact.

Key Applications

Simplified EHL methods for smooth surfaces have interesting applications. They have been used to predict the central film thickness (Christensen 1979), the minimum film thickness, the pressure spike location, and nature of this spike (Greenwood and Morales-Espejel 1995). Similar solutions have been adapted to handle transient effects in lubrication, shear thinning with sliding, and waviness at the inlet. See Morales-Espejel and Wemekamp (2008) for a complete review. Finally, these solutions have proven to be valuable in testing the accuracy of numerical solutions, e.g., (Morales-Espejel et al. 2005). Paradoxically, they become, in general, more accurate as the load in the



Simplified EHL Solution Methods, Fig. 8 Clearance (H) and pressure fluctuations (P) for the asperity geometry for Fig. 7. Solution given by Hooke et al. (2007)

contact increases and/or the lubricant entrainment velocity is reduced (dry contact limit). However, many attempts have proven friction in rolling/sliding EHL contacts difficult to predict with simplified solution methods (in fact, this is the case with any method!). The reasons being mainly the need of non-Newtonian behavior in the “parallel” zone combined with thermal effects to avoid overestimations.

Simplified solutions for rough surfaces have extensive application in studying “real” roughness behavior, since

numerical solutions struggle to accurately describe these transient cases. Simplified solutions have been used to study roughness and indentation effects in rolling contact life (Gabelli et al. 2008) and also in the ranking of surface finishing processes (Morales-Espejel et al. 2003).

Cross-References

- [EHL Film Thickness Behavior](#)
- [EHL Governing Equations](#)

References

- H. Christensen, A simplified model of EHL of rollers. The central and exit solutions. *ASLE Trans.* **22**, 323 (1979)
- A. Gabelli, G.E. Morales-Espejel, E. Ioannides, Particle damage in Hertzian contacts and life ratings of rolling bearings. *STLE Tribol. Trans.* **51**, 428–445 (2008)
- J.A. Greenwood, An extension of the Grubin theory of elastohydrodynamic lubrication. *J. Phys. D: Appl. Phys.* **5**, 2195–2211 (1972)
- J.A. Greenwood, K.L. Johnson, The behaviour of transverse roughness in a sliding elastohydrodynamically lubricated contact. *Wear* **153**, 107–117 (1992)
- J.A. Greenwood, G.E. Morales-Espejel, The behaviour of transverse roughness in EHL contacts. *Proc. Inst. Mech. Eng. Part J* **29**, 121–132 (1994)
- J.A. Greenwood, G.E. Morales-Espejel, Pressure spikes in EHL, in *Lubricants and Lubrication, Proceedings of the 21st Leeds-Lyon Symposium on Tribology* Elsevier Tribology Series, vol. 30 (ed by D. Dowson et al.), (Elsevier, Amsterdam, 1995), pp. 555–564
- A.N. Grubin, Investigation of the contact of machine components, in Kh.F. Ketova (ed.) *Central Scientific Research Institute for Technology and Mechanical Engineering, Moscow (DSIR translation No. 337)*, vol 30, 1949
- C.J. Hooke, Surface roughness modification in elastohydrodynamic line contacts operating in the elastic piezoviscous regime. *Proc. Inst. Mech. Eng. Part J* **212**, 145–162 (1998)
- C.J. Hooke, Y.K. Li, G.E. Morales-Espejel, Rapid calculation of the pressures and clearances in rolling-sliding elastohydrodynamically lubricated contacts. Part 2: general non-sinusoidal roughness. *Proc. Inst. Mech. Eng. Part C* **221**, 555–564 (2007)
- G.E. Morales-Espejel, A.W. Wemekamp, Ertel-Grubin methods in elastohydrodynamic lubrication – a review. *Proc. Inst. Mech. Eng. Part J* **222**, 15–34 (2008)
- G.E. Morales-Espejel, P.M. Lugt, J. van Kuilenburg, J.H. Tripp, Effects of surface micro-geometry on the pressures and internal stresses of pure rolling EHL contacts. *STLE Tribol. Trans.* **46**, 260–272 (2003)
- G.E. Morales-Espejel, M.L. Dumont, P.M. Lugt, A.V. Olver, A limiting solution for the dependence of film thickness on velocity in EHL contacts with very thin films. *STLE Tribol. Trans.* **48**, 317–327 (2005)
- C.H. Venner, Multilevel solutions of the line and point contact problems, in PhD Dissertation, University of Twente, Enschede, The Netherlands, 1991
- C.H. Venner, A.A. Lubrecht, Amplitude reduction of non-isotropic harmonic patterns in circular EHL contacts, under pure rolling, in *Lubrication at the Frontier, Proceedings of the 25th Leeds-Lyon Symposium on Tribology, Elsevier Tribology Series* 36 vol. 34, (ed. D. Dowson et al.), (Elsevier, Amsterdam, 1999), pp. 151–162

- A. von Mohrenstein-Ertel, Die Berechnung der hydrodynamischen Schmierung gekrümmter Oberflächen unter hoher Belastung und Relativbewegung, in *Fortschritts-berichte VDI, Ser. 1, No. 115* eds. by O.R. Lang, P. Oster, (VDI Verlag, Düsseldorf, 1949). First published in 1945 in Russian under the name A.M. Ertel
- P.E. Wolveridge, K.P. Baglin, J.F. Archard, The starved lubrication of cylinders in line contacts. *Proc. Inst. Mech. Eng.* **185**, 1159 (1970–71)

Simulation of Physiological Conditions: An Overview

MARKUS A. WIMMER, MICHEL P. LAURENT
Department of Orthopedics, Rush University Medical Center, Chicago, IL, USA

Synonyms

Testing conditions in biotribology; Testing input; Testing procedures

Definition

This entry provides an overview of realistic testing input and conditions for various testing procedures in biotribology. Although it may be desirable to analyze implants under real conditions, that is, implanted into the human body (in vivo), this is for practical and ethical reasons often undesirable. Hence, preclinical testing relies on realistic simulation models that allow estimating wear properties and durability of prostheses reliably in vitro.

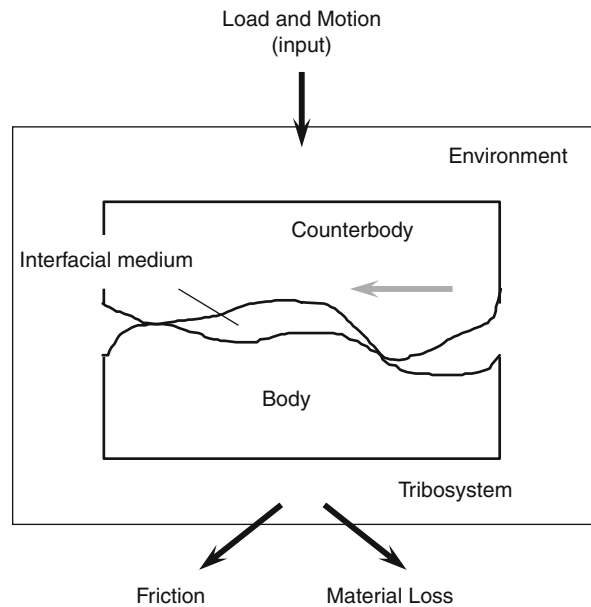
Scientific Fundamentals

Background

In the past, many total joint prostheses have shown serious problems such as component failure and joint loosening due to wear. The clinically unsuccessful experience with Teflon® as replacement material for cartilage at the hip joint has made this very obvious (Charnley 1979). This well-known low friction material had extremely poor wear characteristics when articulating against a metallic counterbody in vivo. Large amounts of generated debris caused severe inflammatory reactions, which rapidly led to loosening of the artificial device. This emphasized the need to consider all the characteristics of a material under tribological stresses carefully, prior to its implantation into the human body.

System Analysis

In order to understand the tribological interactions of an artificial joint in the human body, it may be viewed as



Simulation of Physiological Conditions: An Overview,

Fig. 1 General description of the bio-tribosystem consisting of four principal elements: the two bodies in contact, the interfacial material, and the environment. All these elements can affect each other and change the mechanism of interaction

a tribological system following the approach of Czichos (1978). The two articulating bones comprising the joint are the body and the counterbody. The synovial fluid is the interfacial medium. It maintains a temperature of 37 °C during function and provides lubrication as well as cooling to the artificial devices. Although small in volume (2–4 ml in the larger joints), it is replenished multiple times during daily activity. Environmental parameters (e.g., oxygen pressure, pH value, temperature, etc.) are regulated by the human body and are kept within tight tolerances as long as access is provided (crevices between artificial parts are often problematic). The loads and motions during daily activity determine the input to the system and the moment of friction and the generation of particulate debris and/or metal ions characterize the output (Fig. 1).

System Input

The daily activities of the patient determine the tribological stresses at the joint. There is relatively good knowledge about the daily motion and loading collectives in patients following total hip and total knee arthroplasty (i.e., the replacement of the diseased hip or knee with an artificial joint). There is a limited understanding for the spine, and

information becomes scarce and speculative for most of the other joints of the human body.

A recent study of the frequency and duration of daily activities in patients after total hip arthroplasty has demonstrated that patients spend most of their time sitting (44.3 % of the time), followed by standing (24.5 %), walking (10.2 %), lying down (5.8 %), and stair climbing (0.4 %) (Morlock et al. 2001). Based on this study, the average patient loads the hip more than 3,000 times with walking steps, and over 150 times with stair maneuvers. Resting periods comprise a substantial part of daily activity (Nassutt et al. 2003): 2- to 5-s intervals have been found 100 times per hour; 10- to 30-s intervals of standstills occur up to 26 times per hour. Hence, in vivo joint motion is discontinuous and a permanent fluid film unfeasible. Hence, any materials in contact must tolerate boundary lubrication.

Similar findings have been made for the artificial knee joint. A summary of the daily activity spectrum is given in Table 1 (Orozco et al. 2008). Transitions include chair maneuvers (sitting and rising), comprising about 4 % of the total activity spectrum. It needs to be considered that during a chair and/or stair cycle the knee joint typically undergoes higher tibial loads and motions than during a walking cycle. Since wear is a function of load and motion, chair and stair cycles generate more debris than walking, despite their lower frequency.

Less is known about the spine. Although the kinematic and kinetic parameters of the cervical and lumbar spine have been investigated, and consistent biomechanical range of motion inputs are in widespread use to simulate a variety of daily living activities, the frequencies at which these activities occur are largely unknown. Based on one million yearly gait cycles combined with significant yearly bends in the 100 thousands, as many 100 million “loading cycles” have been predicted over the 40-year lifetime of a lumbar spine implant (Kostuik 1997). However, to date no measurements have been published to support this estimate. Movement frequency data of the cervical spine during various activities are now available (Cobian et al. 2009). This study demonstrated that there are a high numbers of overall motion cycles, which are highest during athletic activity ($8,500 \pm 7,500$ cycles/h) and lowest during sleep (250 ± 200 cycles/h).

Joint Loading

Joint forces have been calculated using mathematical models. However, due to the redundancy of internal muscle and soft tissue structures, an indeterminate problem has to be solved (which introduces uncertainty). Lately, measured forces became available. The most

complete data sets for various implants at the hip, knee, shoulder, and vertebral body replacements are available from Bergmann et al., through the Orthoload database (Bergmann 2011). For example, for the hip joint it is known that the average peak force is approximately 2.4 times the body weight (BW) when walking at a normal speed of 1.1 m/s. This is slightly higher than standing in single leg stance. When going upstairs the joint contact force is 2.5 BW and downstairs 2.6 BW. The peak contact forces are highest during unexpected maneuvers, for example, stumbling, where peak forces reach 8 BW. This is considerably larger than the impact during jogging, which amounts to 5.5 BW.

It should be noted that there is an enormous intra- and inter-variability of forces during daily activities among patients. Therefore, patient-specific modeling to assess individual conditions at the articulation has still its place. This is even more important, since many studies have reported that total joint patients do not achieve a normal walking pattern compared with age-matched control subjects.

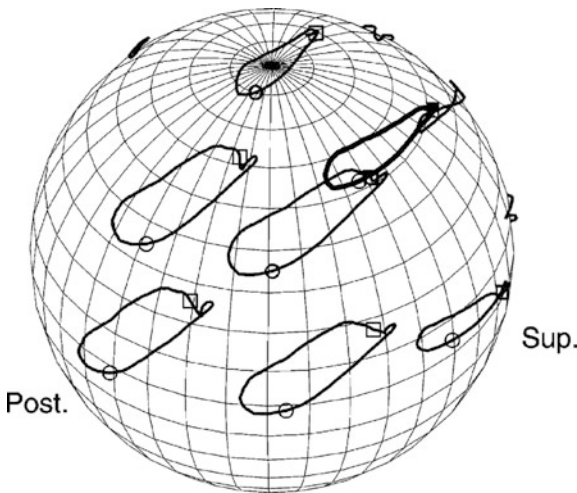
Wear Path

The typical wear path in artificial joints is multidirectional, that is, the movements of the forward and backward stroke do not lie on identical geometrical lines. However, there are time intervals during the gait cycle where the strokes match. Thus, the artificial joint undergoes a mixture of reciprocating and multidirectional sliding. Saikko and Calonius (2002) calculated the hip “slide tracks” between ball and socket for walking based on Euler angles (Fig. 2). Their findings clearly demonstrate that the motion tracks cross each other, leading to multidirectional shear stresses at the surface (which is important for polyethylene wear, see ► [Ultra-High-Molecular-Weight Polyethylene \(UHMWPE\) as a Bearing Material in Hip Joint Replacements](#)). The same group later showed that there are distinct differences between the shapes of the wear tracks generated by contemporary hip simulator designs, which will obviously lead to differences in wear.

Similar observations have been made for other joints, as for instance for the knee. To calculate the contact path during level walking between femoral condyle and tibial plateau, Swanson et al. (2007) defined point clouds of the prosthesis components and aligned those with measurements taken from post-operative radiographs. The femoral component point cloud was then transformed according to six degrees of freedom kinematics measured using gait analysis. The shortest distance between the tibial and femoral point clouds was defined

Simulation of Physiological Conditions: An Overview,
Table 1 Average of activity transition count, locomotion distance and walking speed together with standard deviation, minimum and maximum values for 13 TKR patients. Note that the joint is loaded by every second step

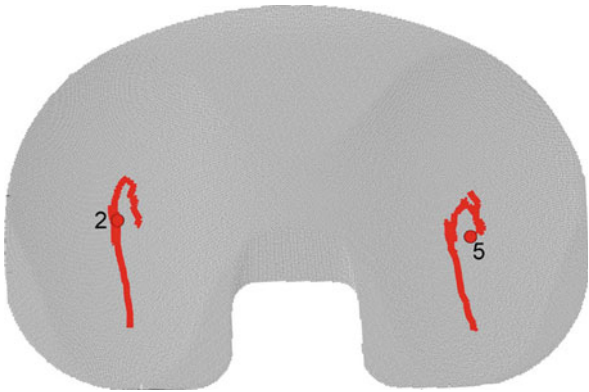
Activity	Count average	St. Deviation	Minimum	Maximum
Walking (steps)	5,857	2,116	3,067	8,461
Running (steps)	7	12	0	38
Stairs (count)	215	165	32	594
Transitions (count)	465	125	245	662
Total Distance (m)	4,005	1,406	2,070	6,430
Walking Speed (m/s)	1.02	0.13	0.83	1.29



Simulation of Physiological Conditions: An Overview,
Fig. 2 Slide tracks of selected points on the femoral head of a hip prosthesis during gait. The track of the force vector is drawn with a *thicker line*. The *square* indicates heel-strike and the *circle* toe-off (Reproduced from Saikko and Calonius 2002)

as the contact point at each instance of stance. [Figure 3](#) shows an example of such a wear path at the tibial plateau.

It needs to be considered that there are a variety of slide tracks during activity of daily living, which include many other activities than walking. Although it is known that the



Simulation of Physiological Conditions: An Overview,
Fig. 3 An example of the tibiofemoral contact pathway of a knee prosthesis during gait. Heel strike is the posterior most point (hollow dot) on both the medial and lateral sides (Reproduced from Lundberg et al. 2010)

shape of the slide track is of fundamental importance with respect to the wear rate, very little specific knowledge about the dependence of wear on different types of multi-directional motion is currently available.

Body and Counterbody

Several material combinations have been tried and are in use for clinical applications (see ► [Modified UHMWPE for the Hip Joint \(Particle Filled and Reinforced\)](#), ► [Self-Mating Metal Articulations in the Hip Joint](#), ► [Self-Mating Ceramic Applications in the Hip Joint](#)). Hence, simulator studies should take the variety into account and consider that different degradation processes may apply. For example, polymers may undergo aging and embrittlement, while metals may suffer combined processes of corrosion and wear. Appropriate environmental conditions during testing may help to identify some of the weaknesses.

In general, one can differentiate between hard-on-soft and hard-on-hard bearing types for biomedical applications. In the case of hard-on-soft, combinations of metal on polymer or ceramic on polymer have been tried. Typically, ultra-high molecular weight polyethylene (UHMWPE) and polyether-ether-ketone (PEEK) are utilized for the soft counterpart. Polyurethanes have also been tried. Metals are made from cobalt-based alloy, nitrated titanium (alloy), and various stainless steels. Typical ceramic representatives are alumina and zirconia. Successful hard-on-hard bearings, which are still in clinical use, comprise of cobalt base metals and alumina ceramics.

Interfacial Medium

Artificial joints are lubricated by neo-synovial fluid. This fluid differs in some biochemical and rheological aspects from healthy synovial fluid due to the preceding osteoarthritic disease process and surgical procedure. It also may contain wear particles and bone chips contributing to third body wear. Ideally, the lubricating film completely separates both articulating elements. This scenario requires a large contact area, sufficiently high relative velocities, and sufficiently smooth surfaces, which is the exception for joint replacements. Theoretically, the combination of a large femoral head (leading to high relative velocity), a small clearance between ball and socket (yielding a large lubricated area), and smooth bearing surfaces could facilitate hydrodynamic lubrication. However, these effects are limited to very few currently available prosthetic geometries, as has been computed by Dowson et al. (1992). In addition, for most practical applications, surface roughening during running-in has to be considered. Hence, boundary lubrication is the dominant lubrication mode that has to be accounted for in artificial joints.

Key Applications

A hierarchical analysis, where the models of wear approximate reality more and more, provides an efficient approach to test the wear and friction properties of new biomaterials. It may start with simple screening tests and end with sophisticated joint simulators.

Screening Tests

There are three screening wear tests that are particularly relevant to orthopedic applications:

1. Pin-on-disk – The simplest; appropriate for basic tribological properties
2. Pin-on-flat – The most prevalent
3. Biaxial pin-on-ball – Intermediate in sophistication, between pin-on-flat and simulator

Pin-On-Disk

This configuration entails a pin subjected to a constant vertical force, sliding on the flat face of a rotating disk, describing a circular, unidirectional path. The main advantage of this configuration is that it offers simple conditions for friction and wear measurements. Guidance for this test is given by ASTM Standard G 99 “Standard Test Method for Wear Testing with a Pin-on-Disk Apparatus.” By measuring the tangential force, the coefficient of friction is easily determined. Although the

tip geometry is typically spherical, rounded and flat tip geometries have also been used. The wear of the pin can often be determined by measuring its loss in height. The wear of the disk can be determined by weight loss or by profilometry. The method is suitable for obtaining friction and wear information on any type of material combination used in orthopedic applications (polymer-on-metal, metal-on-metal, ceramic-on-ceramic, etc.) For a polymer-metal or polymer-ceramic couple, the pin can be chosen to be either of the materials, depending on the information sought. The method has been successfully used to determine the frictional interaction between material couples as a function of lubricant type and composition, however, the lack of cross-shearing due to unidirectional sliding makes it unsuitable for predicting the wear of polymers for orthopedic applications.

Pin-On-Flat

It has been extensively used to screen polymeric materials sliding against metal, but may also be used for hard-on-hard combinations. The test configuration entails the end of a cylindrical pin sliding against a flat counter-face. The end of the pin may be flat or rounded. The typical configuration for testing polymers for use in joint prostheses consists of a flat-faced cylindrical pin sliding against a flat metal counter-face. Because of its importance, this test has been standardized with ASTM Standard F 732, “Wear Testing of Polymeric Materials Used in Total Joint Prostheses.” This standard specifies three variants of the test: (1) for linear reciprocation wear motion applications, such as hinged knees; (2) for hip-type motion; and (3) for linear motion delamination wear applications, mainly applicable to incongruent metal-polymer contact as encountered in knee prostheses. Variant 2 emulates the multidirectional motion found at the bearing surfaces of hip prostheses, determined to be essential to properly evaluate polyethylene for biomedical applications.

More complex wear screening tests of the pin-on-flat type have recently become available (e.g., Saikko and Kostamo 2011). These devices can be programmed to produce virtually any slide track shape and load profile. It has been shown that random motion resulted in a higher wear factor than circular motion for polyethylene despite the same sliding distance. However, the type of load, random vs. static, proved unimportant in this investigation. The principal advantage of using a random track is that possible unrealistic wear phenomena related to the use of fixed track shapes can be avoided.

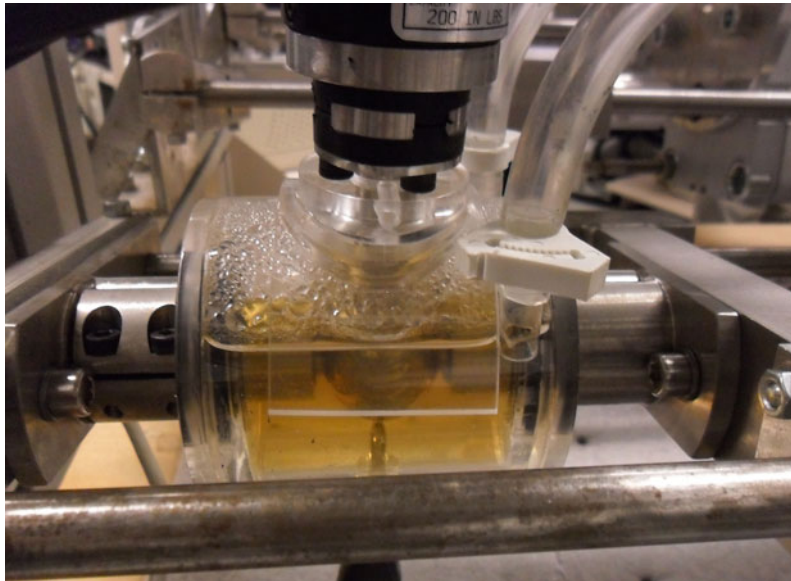
Pin-On-Ball

Intermediate in sophistication between pin-on-flat and hip simulator tests, this test was explicitly conceived as a method to screen and analyze bearing surfaces used in total hip replacement (Wimmer et al. 1998). It entails two cylindrical pins and a ball oscillating rotationally about perpendicular axes (Fig. 4). The concave end of the pin produces a conforming, equatorial contact on the ball. If the axes of the pins and ball are run in phase difference, the resulting biaxial motion yields wear tracks with crossing paths that approximate wear tracks observed *in vivo*. If they are run without phase difference, nearly linear motion tracks are achieved. This flexibility allows evaluation of the impact of motion trajectory on the wear of different candidate materials. The load is applied along the pin axis and can be kept constant or varied cyclically to simulate loads during daily activity. Arbitrary combinations of materials can be tested, such as soft-on-hard (e.g., a UHMWPE pin against a Co-Cr-Mo ball) and hard-on-hard (e.g., metal against metal, ceramic against ceramic). The apparatus is particularly well suited to test coatings on the ball surface since they are applied on original parts. The pin-on-ball assembly is immersed in a chamber containing the lubricant, such as diluted bovine calf serum. The friction between pin and ball is determined from a torque sensor, which sits at the rotational axis of the ball. Linear wear and deformation

are measured using an LVDT displacement sensor aligned with the pin. For UHMWPE, the effect of creep and swelling may be reduced by loading and soaking the wear couple for a predetermined time prior to starting the wear test. Wear may also be determined gravimetrically. No standards are currently associated with this biaxial pin-on-ball test.

Simulator Tests

As useful the screening tests may be, ultimately it is essential that wear tests be performed using the prosthetic components themselves in a manner that simulates as closely as possible the relevant physiological conditions *in vivo*. This is typically achieved by a duplication of the physiological motions, namely, flexion-extension, adduction-abduction, and internal-external rotations during activities of daily living. In today's simulators, these activities are usually restricted to gait only. In the future, it will be necessary that a variety of profiles can be inputted to simulate additional activities other than walking (e.g., running, stair climbing, stair descending, and seating maneuvers, etc). Despite their low frequency, these activities may have a considerable impact on wear cumulative load as described above. Further, a simulator should permit anatomic positioning of the joint (e.g., cup above head in case for the hip joint), and allow the utilization of lubricants that mimic neo-synovial fluid.



Simulation of Physiological Conditions: An Overview, Fig. 4 Photograph of a single station of the pin-on-ball testing apparatus. Note the two pins that press equatorially onto the ball. Pins and ball are immersed in diluted bovine serum, which serves as the testing lubricant

Hence, the test chambers need to be constructed of materials, which are inert to the lubricant and are sealed to prevent fluid evaporation and the ingress of contaminants. In order to simulate several years of the life-span of prostheses within a few months, simulators must run unattended 24 h a day, 7 days a week, except for occasional checks and periodic processing of the specimens for cleaning and measuring wear.

It is important that the simulator reproduces the wear mechanisms observed in vivo as closely as possible. Hence, it is necessary that the following applies between in vivo and in vitro observation:

1. The magnitude of the wear rates and ranking of materials are similar
2. The microscopic appearance of the wear surfaces is alike
3. The debris morphology and size distribution is comparable.
4. The microstructural changes at the surfaces are comparable

Details about simulator testing of the hip, knee, spine, and other joints and their documentation in ISO and ASTM standards are discussed in the following entries:

► [Testing of Artificial Hip Joints](#), ► [Testing of Artificial Knee Joints](#), and ► [Testing of Artificial Discs](#).

Cross-References

- [Self-mating Ceramic Applications in the Hip Joint](#)
- [Self-mating Metal Articulations in the Hip Joint](#)
- [Testing of Artificial Hip Joints](#)
- [Testing of Artificial Knee Joints](#)

References

- G. Bergmann, Loading of orthopaedic implants, 2011, www.orthoload.com, Accessed on 20 Nov 2011
- J. Charnley, *Low Friction Arthroplasty of the Hip* (Springer, Berlin, 1979)
- D. Cobian et al., Task-specific frequencies of neck motion measured in healthy young adults over a five-day period. *Spine* **34**, 202 (2009)
- H. Czichos, *Tribology – A Systems Approach to the Science and Technology of Friction, Lubrication and Wear* (Elsevier, Amsterdam, 1978)
- D. Dowson, Friction and wear of medical implants and prosthetic devices. *ASM Handbook. Friction, Lubrication, and Wear Technology*. vol. 18. American Society for Metals, (1992), ASM International, Materials Park (OH), p. 656
- J. Kostuik, Intervertebral disc replacement. Experimental study. *Clin. Orthop.* **337**, 27 (1997)
- H. Lundberg, M. Wimmer, The effect of tibiofemoral contact path centroid location on TKR contact forces. In *Proceedings of the ASME Summer Bioengineering Conference*, 2010, ASME, New York, SBC2010-19407
- M. Morlock et al., Duration and frequency of every day activities in total hip patients. *J. Biomech.* **34**, 873 (2001)
- R. Nassutt, M. Wimmer et al., The influence of resting periods on friction in the artificial hip. *Clin. Orthop.* **407**, 127 (2003)
- D. Orozco et al., Occurrence of daily activity transitions in an active TKR population. *Trans. Orthop. Res. Soc.* **33**, 1975 (2008)
- V. Saikko, O. Caloni, Slide track analysis of the relative motion between femoral head and acetabular cup in walking and in hip simulators. *J. Biomech.* **35**, 455 (2002)
- V. Saikko, J. Kostamo, Random POD – a new method and device for advanced wear simulation of orthopaedic biomaterials. *J. Biomech.* **44**, 810 (2011)
- A. Swanson et al., Analysis of the tibio-femoral contact point in total knee replacement using a marker based motion analysis system. In *Proceedings of the ASME Summer Bioengineering Conference*, 2007, ASME, New York, SBC2007-176757
- M. Wimmer et al., A new screening method designed for wear analysis of bearing surfaces used in total hip arthroplasty. *Alternative Bearing Surfaces in Total Joint Replacement*, ed. by J. Jacobs et al., ASTM STP 1346, 1998, ASTM, West Conshohocken, PA, p. 30

Simulation of the Tribological Behavior of Engine Components

GARY C. BARBER, MARK J. MALATESTA

Department of Mechanical Engineering,
Automotive Tribology Center, Oakland University Dodge
Hall of Engineering, Rochester, MI, USA

Synonyms

[Bench testing of engine components](#)

Definition

Bench testing of engine components run under operating conditions that duplicate the tribological behavior of components run in fired engines.

Scientific Fundamentals

Engines are constantly being subjected to demands for enhanced performance. These include increased horsepower to weight ratios, reduced fuel consumption, and reduced oil consumption. Additional factors that affect engine design are imposed by social and political considerations. Social legislation has resulted in limits to particulate matter that can be discharged to the environment. Alternatives for strategic materials must be available without adversely affecting performance. These factors pose significant challenges to engine design and development.

The environment to which engine components are exposed is not fully understood. Large differences exist between various engines of the same type and even within a single engine where cylinder environments are expected

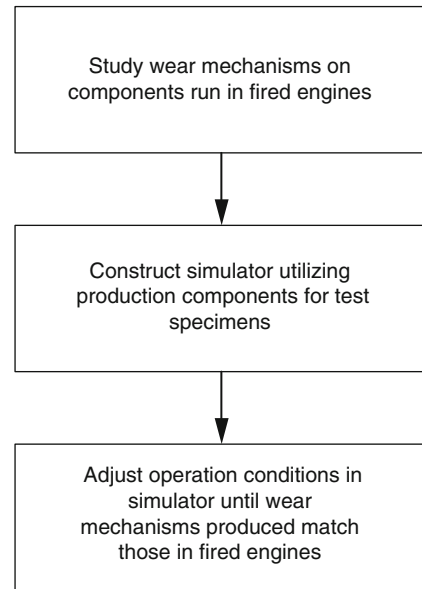
to be identical. The necessity of developing an understanding of an engine's environment is demonstrated by the fact that inherent engine variables may cause a wide variation in component life, which in some cases is more significant than changes in materials or component design.

The accepted measure of an engine component's performance comes from engine tests and ultimately from field use. The value of such tests derives not from the fact that all variables present are known, but, whether recognized or not, these variables comprise the system in which the component operates. Why a particular component performs well in an engine may not be known, but that it did or did not survive is easily determined.

Design iterations for engine components are based primarily on the results of engine tests. These methods will continue to be necessary and useful. However, the degree to which they are utilized will decline because engine tests are extremely expensive. More economical, controllable, and efficient test methods are being sought. Hence, there is a need for accelerated testing techniques to replicate engine component wear and failure mechanisms. A complete understanding of wear mechanisms and a means to isolate various material, geometric, dynamic, and environmental parameters are necessary. Through isolation, consistent testing can be accomplished and variation of parameters will allow designers to understand the interdependence of the various factors. The use of economical test methods is needed to simulate the tribological performance of engine components.

The primary requirement of a simulator is to rank various component designs and materials in the same order as a field test. The concept of a simulator would be violated by strict attention to the simulation of every detail of an engine. The model or simulator should provide a controllable and repeatable behavior within a reasonable time. Only the duplication of those parameters that yield wear comparable to that obtained in service is necessary. The simulation of all parameters is unreasonable and adds unnecessary complexity. A simulator should offer an alternative to the method of consumer feedback. It must never replace it; rather it should shorten the design process and enhance product improvement.

When designing a bench test to simulate the tribological behavior of engine components it is desirable to duplicate as many of the geometric, material, and operating conditions, within economic constraints, as are observed in an operating engine. The geometry and material are typically duplicated by constructing a simulator that utilizes actual engine components. However, since all of the operating conditions are not duplicated in the



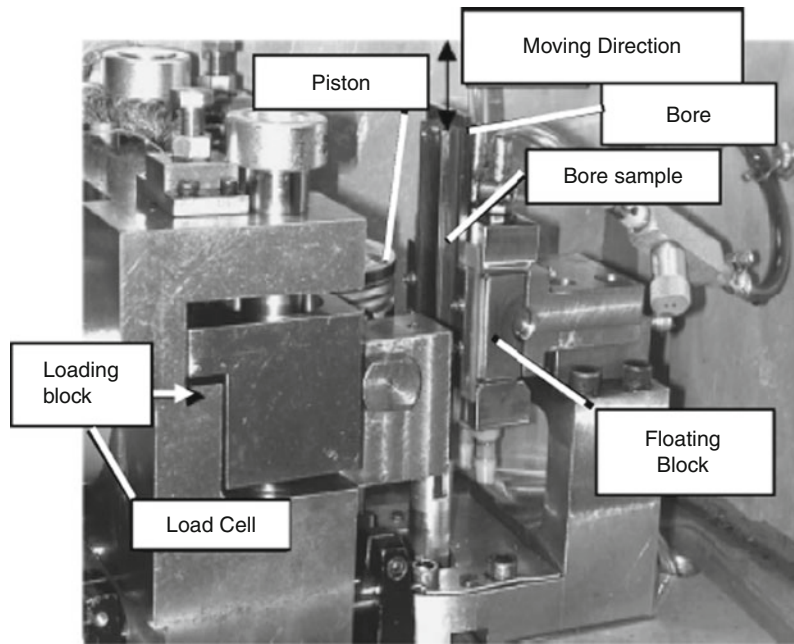
Simulation of the Tribological Behavior of Engine Components, Fig. 1 Tribological simulation process

simulator, there will always be some doubt as to whether the simulator is producing valid results. One way to help ensure validity of the simulator is to select operating conditions for the simulator that produce the same wear mechanisms in the simulator as are produced in fired engines (Barber and Ludema 1987). This process involves studying wear mechanisms produced in fired engines, constructing the simulator, and adjusting the operating conditions of the simulator until the wear mechanisms produced in the simulator match those observed in fired engines, see Fig. 1.

Key Applications

Engine Cylinder Bore Tribology

The piston rings, rather than the piston, are the main cause of cylinder wall wear, and cylinder wall wear is typically greatest at the highest point of piston ring travel. There is higher wear at this point since this is the location of highest contact pressure and the location of the thinnest oil film. Cylinder wall wear can occur by several mechanisms, with abrasive wear and corrosive wear the most commonly observed mechanisms. Catastrophic damage, i.e., scuffing, can also occur on cylinder walls. Scuffing may result from high operating temperatures, high loads, or interruptions of oil supply. Scuffing results in considerable roughening of the cylinder wall and, if allowed to



Simulation of the Tribological Behavior of Engine Components, Fig. 2 Cylinder wall tribology simulator

continue, seizure between the piston/piston ring and the cylinder wall may result.

The cylinder bore (wall) is the most commonly simulated engine component, with various simulators constructed to study the tribological behavior of piston rings or pistons sliding against segments of cylinder bores (Williams and Daniel 1955; Tung and Tseregounis 2000). One example is the test machine shown in Fig. 2, which simulates the contact between a piston skirt or piston rings sliding against a cylinder bore (Gupte et al. 2008). A single-cylinder engine driven by a variable-speed motor provides a reciprocating motion for the cylinder bore sample. A horizontal load is applied by an actuator that forces a piston or piston rings against a bore sample through a loading block. A horizontal pin is floated in both pin bores of the piston. The piston can rotate and slide along the pin so that it self-aligns with the bore sample's curvature.

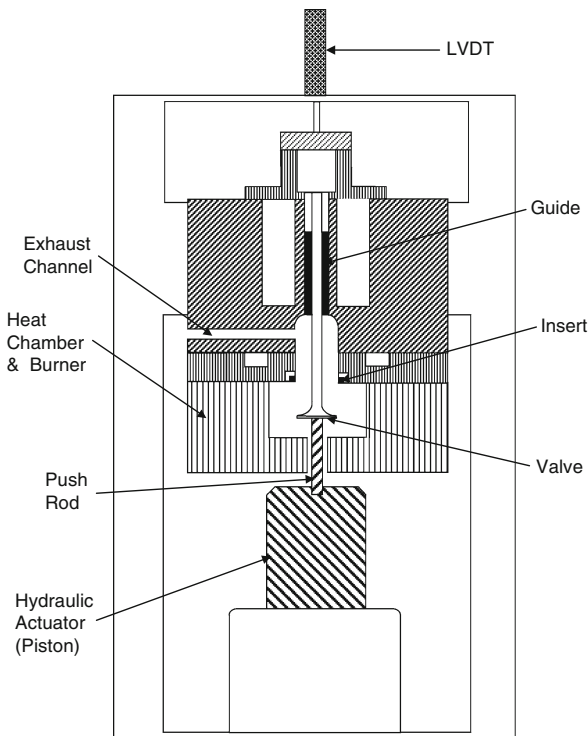
The normal load, the frictional force, and the temperature of the contact surfaces are measured during testing. The normal load is measured by a load cell located between the loading block and the actuator. When a cylinder bore reciprocates against a piston (or ring) sample, the frictional force between the two samples is measured by a second load cell, which is located beneath the piston. A thermocouple installed on the inner side of the piston just beneath the sliding portion of the piston

(or ring segment) is used to measure temperature during the test. The output from the load cells and thermocouple are converted into digital signals and then input to a computer.

The test machine can be used for wear studies or scuffing studies. When used to study scuffing, the normal load is increased at constant time increments until scuffing occurs. Other typical operating conditions related to the use of this simulator to study scuffing are as follows (Gupte et al. 2008): reciprocating speed of the cylinder bore sample of 25 Hz (1,500 rpm), stroke length of 112 mm and SAE 5W-30 oil. The tests are monitored using a computer to collect real-time data and stopped when scuffing occurs. Scuffing is detected by sharp increases in temperature and friction force. The scuffing load is defined as the magnitude of the normal load when scuffing occurs and is used to rank scuffing resistance of materials and piston/piston ring/cylinder bore designs.

Engine Valve Tribology

Engine valves are exposed to a particularly harsh environment consisting of high temperatures and corrosive products of combustions. This harsh environment contributes to valve seat wear, which can occur by several mechanisms. These include adhesive wear, abrasive wear, corrosive wear, and surface fatigue. The simulator shown in Fig. 3 has been shown to effectively duplicate



Simulation of the Tribological Behavior of Engine Components, Fig. 3 Schematic of seat wear test fixture

the mechanisms of adhesive wear and abrasive wear (Malatesta et al. 1993).

This test rig is capable of handling a wide variety of valve sizes, ranging from small passenger car valves to large, heavy duty valves. This allows the test to be applied to both light and heavy duty engine applications. It utilizes full-size valves and inserts to differentiate between samples.

The machine uses an actuator to seat the valve against its insert. To simulate the loading of the valve due to combustion pressure, a small diameter rod is positioned in a cup atop the hydraulic ram and held in a dimple in the face of the valve. The loading that results from this configuration deforms the valve in a manner similar to combustion loading in an engine.

The environmental system of the simulator includes a burner and its exhaust shroud (in Fig. 3 this is represented as the heat chamber) as well as coolant channels. The burner consists of a ring that encircles the face of the valve. By utilizing a propane/oxygen mix as fuel, it maintains the valve at temperatures typical of an internal combustion engine. The body of the heat chamber surrounds the burner and exhausts a large

percentage of the hot gases that have been generated by the system. A secondary exhaust channel is situated above the valve. This is important because gases exhausted from this channel are drawn through the interface between the valve and the insert, as occurs in an actual engine.

Similar to an engine block, cooling channels are used to enhance heat transfer. The primary location of heat transfer occurs between the valve seat and the insert during the seating event. Therefore, a cooling channel is provided in the retainer plate. A second path of heat transfer exists at the interface between the valve stem and the guide insert. The test machine is used to compare wear resistance of various valve materials and geometries.

In summary, simulation of the tribological behavior of engine components is an important process to speed up the engine design process and help ensure optimum performance of engines. Two important examples are the simulation of piston/piston ring/cylinder wall tribology and valve seat/insert tribology. Confidence in the validity of results produced by an engine component simulator can be enhanced by ensuring that the wear mechanisms produced in the simulator match those produced in fired engines.

Cross-References

- [Engine Lubricants](#)
- [Wear Maps](#)

References

- G.C. Barber, K.C. Ludema, The break-in stage of cylinder wear: a correlation between fired engines and a laboratory simulator. *J. Wear* **118**, 57–75 (1987)
- P.S. Gupta, Y. Wang, W. Miller, G.C. Barber, C. Yao, B. Zhao, Q. Zou, A study of torn and folded metal (TFM) on honed cylinder bore surfaces. *Tribol. Trans.* **51**(6), 784 (2008)
- M.J. Malatesta, G.C. Barber, J.M. Larson, S.L. Narasimhan, Development of a laboratory bench test to simulate seat wear of engine poppet valves. *Tribol. Trans.* **36**(3), 627 (1993)
- S.C. Tung, S.I. Tseregounis, *An Investigation of Tribological Characteristics of Engine Oils Using a Reciprocating Bench Test*, SAE technical paper, 2000-01-1781 (2000)
- K.R. Williams, S.G. Daniel, The running-in of engines: choice of cylinder bore finish, in *Proceeding of the Institute of Mechanical Engineers, Automobile Division* (London, England, 1955)

Sintered Oil-Filled Bearings

- [Porous Metal Journal Bearings](#)

Size Effects in Machining Tribology

SHREYES N. MELKOTE¹, SATHYAN SUBBIAH²

¹George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

²Nanyang Technological University, Singapore, Singapore

Definition

Interesting trends in specific cutting energy are observed in metal machining, where new surfaces are created by the relative motion between a cutting tool and workpiece, when certain dimensions (size) associated with the material removal process are reduced to the micron scale. These trends require explanations beyond the commonly accepted assumptions of machining and are referred to as the size effect in machining.

Scientific Fundamentals

Figure 1a shows a schematic of the simplest mode of machining, termed orthogonal machining, where the radiused tool cutting edge is perpendicular to the direction of workpiece motion. The process is essentially removing a layer of thickness t_o from the work material. The parameters related to size effects observed in machining are the uncut chip thickness, t_o , the cutting edge radius, r (Fig. 1a), and the specific cutting energy, u . The specific cutting energy is defined as the energy needed to remove a unit volume of material and, for orthogonal machining, is calculated as,

$$u = \frac{F_c V}{V b t_o}$$

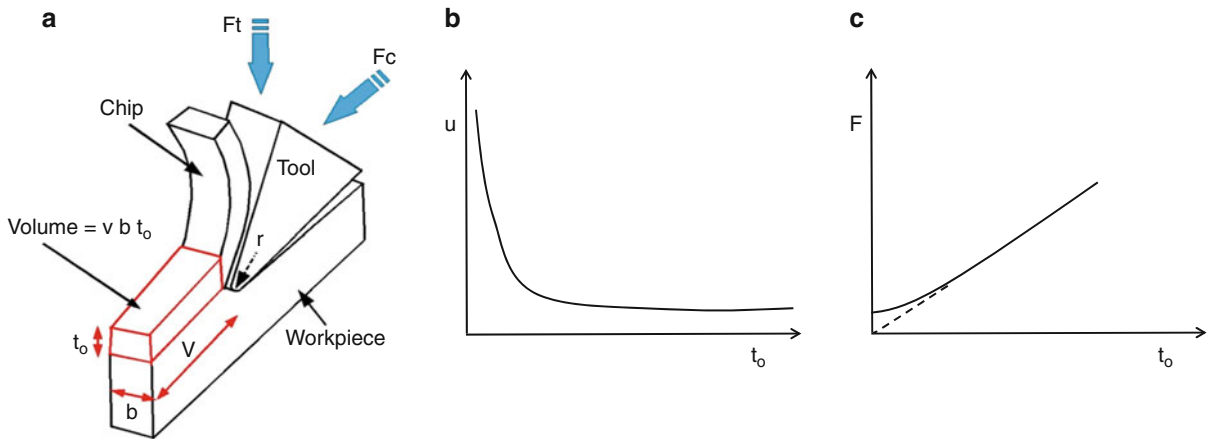
Two forces are primarily involved in orthogonal cutting: the cutting force, F_c , which is in the direction of the cutting velocity, and the thrust force, F_t , which acts normal to the machined surface and the cutting velocity. Factors that cause these components to be higher than normally expected and lead to higher energy input into the machining process can be expected to adversely affect the machined surface integrity. The size effects in machining are observed when the thickness of the machined layer, t_o , formally known as the uncut chip thickness, is gradually reduced, i.e., when thinner and smaller amounts of material are removed. In practice, such a situation is seen in two cases: (a) when a high surface finish requirement dictates smaller and smaller amounts of material removal and (b) when machining is used to create micrometer-sized features using miniature cutting tools.

Observations in Forces and Energy

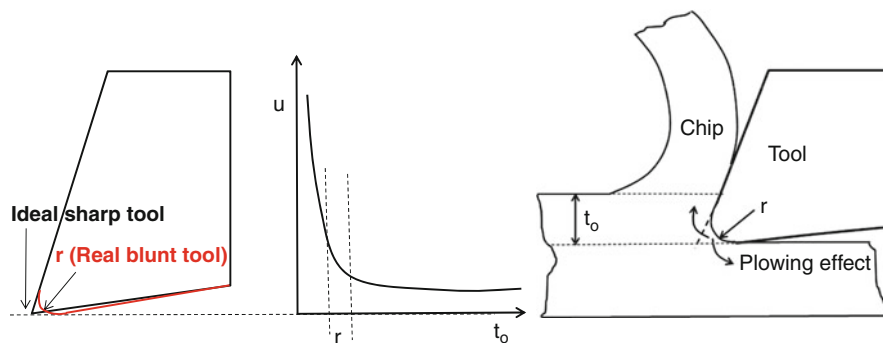
The study of size effect in machining seems to have started with observations of interesting trends in the measured forces and computed energy. Of the several process responses that can be made to study and understand the machining process, the measurement of forces and the associated specific cutting energy are of high importance. The specific cutting energy, u , is seen to increase dramatically as t_o is reduced (Fig. 1b). This was first reported in the 1950s (Backer et al. 1952). The cutting and thrust forces are also seen to decrease disproportionately as t_o is reduced (Fig. 1c). Such observations have been reported in many engineering materials (steel, brass, PMMA, CaF₂, germanium, Al-alloys), under many different machining conditions (V ranging from 0.1 to 200 m/min; t_o ranging from 10 nm to 300 μ m) and using tools of varying edge radii (r ranging from 65 nm to 4 μ m) (Furukawa and Moronuki 1988; Lucca et al. 1991; Liu and Melkote 2006a). These observations indicate that more force and hence more energy is needed to remove material at smaller length scales than previously thought; in other words, one cannot simply extrapolate from the macro to the micro scale. Explanations of this trend in force and energy calculations have led to re-evaluation of several assumptions in traditional cutting models. For example, classical models of machining assume that the material is homogeneous, the tool is ideally sharp, material removal is through shear, fracture is absent, and the apparent tool-chip friction coefficient is constant. All of these assumptions have come under increasing scrutiny when trying to explain the forces and energy trends observed in material removal at the small (micrometer and lower) scale.

Cutting Edge Radius

The effect of cutting edge radius on the machining process at small values of t_o is known to give rise to an important size effect. Cutting tools are never ideally sharp and hence the cutting edge formed at the intersection of the tool rake and flank faces is never an ideal straight line or curve (the rake face is the face over which the chip flows, while the flank face is the relief face of the tool closest to the machined surface). There is always a small radius present at the cutting edge. Studies have shown that the force and energy trends change in the region where the uncut chip thickness becomes comparable to the cutting edge radius, even when the tool is extremely sharp with edge radius in the few tens of nanometers (Lucca et al. 1991; Ng et al. 2006). A tool with a radiused cutting edge is essentially a blunt tool and this causes some pushing of the material immediately ahead of the cutting edge into the chip and into the machined work surface and possibly to the sides.



Size Effects in Machining Tribology, Fig. 1 (a) Orthogonal machining geometry indicating the uncut chip thickness, t_o , the cutting edge radius, r , and the volume of material removed in unit time; note that the cutting edge is perpendicular to the direction of cutting velocity V in orthogonal machining. Also shown are the cutting (F_c) and thrust force (F_t) directions. The workpiece is moving with a velocity V and the width of the work material is b . (b, c) Typical trends in specific cutting energy and forces with reduction in uncut chip thickness. The specific cutting energy is seen to increase significantly while the machining forces tend to level off



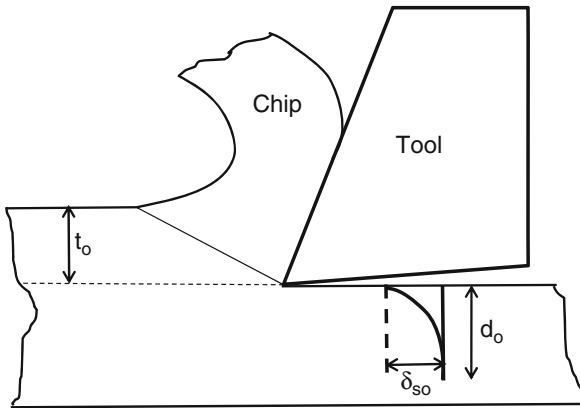
Size Effects in Machining Tribology, Fig. 2 An ideal tool is sharp with zero tool edge radius, while real tools have a finite edge radius, r . Experiments show that the increase in specific cutting energy is seen to start in the region where t_o approaches r . When t_o is comparable to r the plowing effect on the surface integrity is more significant

This is commonly referred to as “plowing” (Fig. 2). A blunt tool also results in a large negative effective rake angle, which tends to push the work material (Komanduri 1971). This leads to higher forces and energy to be expended in the cutting zone and also in turn affects the integrity of the machined surface.

Sub-surface Plastic Flow

Plastic flow occurring in the machined sub-surface does not decrease proportionally at smaller scales of machining. One of the side effects of machining is the alteration of the material surface and sub-surface properties since chip

removal involves considerable force and deformation. This sub-surface plastic flow or “damage” can be characterized by two parameters, the surface plastic flow, δ_{so} , and the depth of damage, d_o (Fig. 3). Analysis of the natural micro-structural flow observed below the machined surface reveals that the sub-surface plastic deformation does not proportionally reduce as the uncut thickness is decreased (Fig. 3) (Nakayama and Tamura 1968; Abdelmoneim and Scrutton 1974). In other words, removal of even small amounts of material results in considerable sub-surface plastic flow. One can also scribe lines on the sides of the work surface and study how the



Size Effects in Machining Tribology, Fig. 3 The sub-surface plastic flow does not decrease proportionally with decrease in t_o . An originally straight line becomes curved, as shown. The depth of the damage (d_o) and the surface plastic displacement (δ_{so}) can be characterized either by the natural flow of the microstructure, as observed in an etched cross section, or by scribing lines on the cross section and observing them before and after machining

lines deform upon machining. Such a controlled study at different edge radii and uncut chip thicknesses show that, while the depth of plastic flow-induced damage scales proportionally with t_o , the surface plastic flow does not. The depth of damage can be expressed as a function of the resultant force, F , and the shear stress, K of the material (Abdelmoneim and Scrutton 1974):

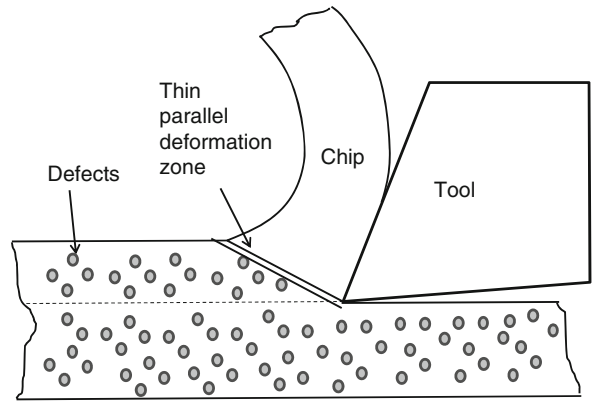
$$d_o = \frac{F}{3K}$$

Material Effects

Machining involves large deformation of the work material to be removed in the form of chips. Strains are of the order of 3–5 and strain rates are of the order of 10^6 s^{-1} . Such large deformation can be accompanied by high temperatures ($\sim 0.5 T_m$). Hence, the material's strength, strain hardening, strain-rate hardening, thermal softening, and fracture properties play an important role in the material removal process and in determining the integrity of the machined surface. The specific material effects that start to become important at smaller length scales of cutting are discussed briefly in the following sections.

Material Strengthening

A consequence of the reduced material removal (smaller t_o) is that deformation happens in a much smaller volume of material and hence the process is influenced by changes



Size Effects in Machining Tribology, Fig. 4 At smaller scales of machining the chance of encountering a defect is less

in material strength at these small scales. Plastic deformation in materials usually starts in areas where defects (e.g., dislocations) are present and hence the presence of defects lowers the material strength. At smaller volumes of deformation the chances of encountering a defect are less (Fig. 4) and this can cause an apparent increase in the material strength thus affecting the forces and energy and consequently the machined surface integrity (Shaw 1950). Most of the deformation in machining occurs in an idealized narrow parallel zone ahead of the tool; this zone becomes even thinner at smaller uncut chip thickness values. Hence, chances of encountering a defect in this small zone are reduced.

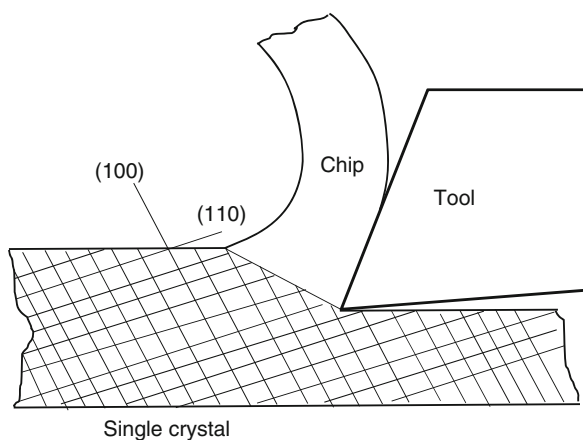
A material's plastic deformation response is normally dependent on the level of strain, strain rate, and temperature. Any changes in these three factors in machining can impact the material strength and hence on the forces, energy, and surface integrity. Experimental data for low carbon steels has shown that as the uncut chip thickness decreases the strain rate in the deformation zone is higher leading effectively to a higher strength of the material (Larsen-Basse and Oxley 1973). Also, the temperature in the secondary deformation zone (tool-chip interface) decreases with a decrease in the uncut chip thickness leading to a reduced thermal softening effect (Kopalinsky and Oxley 1984). This also leads to strengthening of the material and can explain the higher forces and energy expended at smaller scales of machining. Higher forces and energy leave a more severe foot print on the machined surface and sub-surface.

Materials also tend to behave stronger when the deformation is concentrated in a very narrow zone leading to large spatial gradients in strain. Such a situation is seen in machining where deformation is concentrated in

narrow parallel shear planes and the zone becomes even smaller at smaller uncut chip thickness values. Hence, the gradient of strain can be expected to affect the machining process and the surface generated (Dinesh et al. 2001). The effects of strain gradients can be incorporated into plasticity constitutive equations and incorporated into machining simulation models, both analytical and numerical (Joshi and Melkote 2004; Liu and Melkote 2006a). Numerical simulations for Al5083-H116 aluminum alloy have shown that strain gradient strengthening contributes significantly to the force and energy at low cutting speeds (less than 10 m/min) and small uncut chip thickness (less than 10 μm).

Crystalline Orientations

At smaller scales of material removal, the averaging affect of various crystalline grain orientations normally seen in a polycrystalline material is no longer present. The moving cutting edge now encounters individual grains and crystallographic planes in the grains (Fig. 5) and since the material property changes with the crystal orientations this affects the forces, energy, and surface finish (Cohen et al. 1981). The effects of grain level anisotropy on the machining response has been experimentally observed and studied. The chips and surfaces generated in machining of single crystal copper show that the chip thickness and shear angle vary considerably with change in crystallographic orientations (Lee et al. 2000). Surface finish is also affected when the cutting direction relative to the crystal orientation is changed. In addition, orthogonal micromachining of single-crystal aluminum has shown that the machining forces vary with crystallographic orientation (Lawson et al. 2008). This study also reports that



Size Effects in Machining Tribology, Fig. 5 Crystal plane and orientations affect the process at smaller t_o

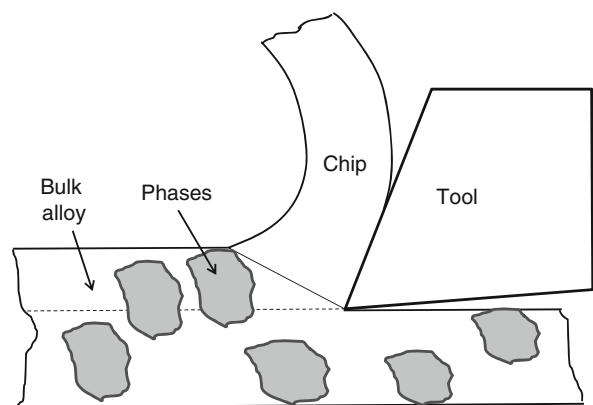
the uncut chip thickness has an orientation-dependent effect on the machining forces. In addition, the chips formed in single crystal aluminum have different morphology and displayed shear-front lamella, the periodicity of which varies with crystallographic orientations and uncut chip thickness.

Phases

Engineering alloys usually have several phases present in them and the phases have different material properties from the bulk alloy. In conventional scale machining, the effects of these phases average out. At smaller scales of machining the tool cutting edge now individually encounters these phases as it moves from cutting one phase to another (Fig. 6). This causes variations in forces, energy, and surface integrity. These variations have been studied in detail (Vogler et al. 2003). For example, in micro-milling of cast iron it is shown that frequencies in the cutting force signal higher than those attributable to the process kinematics can be explained by explicitly considering the interaction between the tool and the multiple phases present in the material. Experiments performed on two compositions of ductile iron phases – pure ferrite and pearlite – show that the nature of variation in the ductile iron cutting force signals can be attributed to the mixture of the phases. In addition, simulation studies show that the frequency of variation is related to the spacing of the secondary (ferrite) phase and the variation magnitude is determined by the size of the secondary (ferrite) phase particle.

Importance of Fracture

In machining, the energy expended to create a new surface via material separation and hence the fracture properties



Size Effects in Machining Tribology, Fig. 6 Cutting through a heterogeneous material with phases at small t_o

of the material can become important, especially at smaller cutting scales. The machining process involves separation of the chip from the work material leading to new surfaces being formed. From this point of view, machining is similar to fracture where crack propagation generates two new surfaces. At conventional length scales, the plastic deformation energy (E_{plastic}) and the frictional energy dissipated (E_{friction}) as the chip moves over the tool rake face form a large portion of the total energy and hence dwarf the energy consumed in material separation (E_{surface}). Also, while cracks and fracture are more readily observed when machining brittle materials, it has been hard to find evidence of fracture leading to material separation in machining more ductile materials. Hence, the fracture component was historically ignored for the large part of machining research. However, at smaller scales of machining the total energy used in plastic deformation and friction energy become smaller and hence the energy for surface formation becomes significant. Traditional machining models have been modified (Table 1) to include this fracture energy component (Atkins 2003). Also, physical evidence of fracture and ductile tearing ahead of the cutting tool edge has been documented recently (Subbiah and Melkote 2007). The presence of the fracture energy component has been shown to account for the observed size effect in force and specific cutting energy as the uncut chip thickness is reduced.

Ductile Regime Machining

Under certain conditions a smooth surface finish is possible in machining of brittle materials. It was first observed in single abrasive grit scratch tests on brittle materials that under certain conditions of load and depth no cracks were observed on the surface of the scratch. The idea was then extended to machining where a critical t_o (Fig. 7) was identified; when machining is performed below this critical t_o crack-free machined surfaces were obtained on brittle materials such as silicon and germanium (Blake and Scattergood 1990). In order to obtain consistent and uniform crack-free surfaces, machining has to be performed on special equipment such as an ultra-precision lathe that maintains good position control with high rigidity and low vibrations. As shown in Fig. 7, the material removal mechanism changes from a ductile

crack-free mode to a brittle mode as the uncut chip thickness value reaches a critical value, t_{oc} . The ability to create crack-free surfaces in brittle materials below certain t_o is commonly referred to as ductile regime machining. The critical uncut chip thickness depends on the material properties; for example, it is higher for silicon than germanium. Parameters such as the tool rake and clearance angles also affect t_{oc} while other parameters such as the surface cutting speed have a negligible effect. Recently, Venkatachalam et al. (2009) have shown that the cracks are absent when the shear stress in the chip formation region is greater than the critical shear stress for chip formation and the stress intensity factor is less than the fracture toughness of the material. The point of transition is said to take place when the fracture toughness is equal to the stress intensity factor.

Friction Coefficient

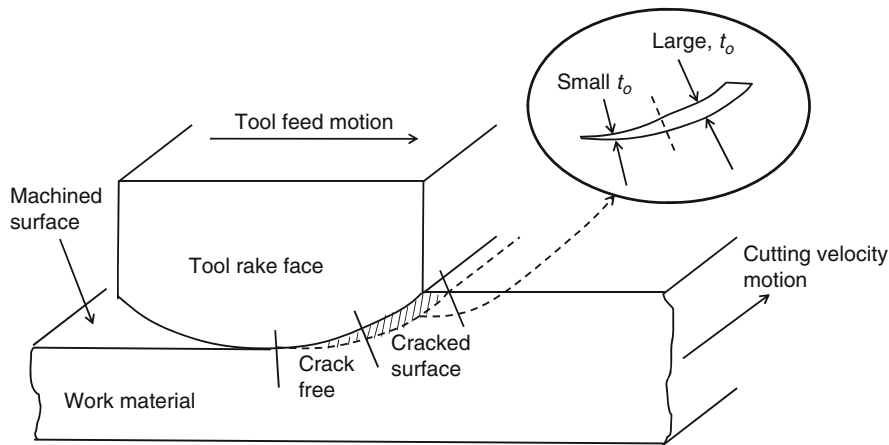
In machining, the chip slides over the tool rake face, which is subjected to considerable normal and frictional forces. This causes the apparent tool-chip friction coefficient to be high and leads to energy dissipation due to the rubbing action. The frictional energy dissipated at the tool-chip interface is a major contributor to the total energy, being second only to the plastic deformation energy dissipated in the primary shear zone. With a decrease in the uncut chip thickness, the mean coefficient of friction at the tool-chip contact increases nonlinearly (Ng et al. 2006). As the uncut chip thickness decreases, the cutting temperature at the tool–chip interface decreases. This leads to an increase in the shear yield strength of the work material at the tool–chip interface. The increase in the shear yield strength of the work material at the tool–chip interface tends to increase the friction coefficient. A similar trend in friction coefficient can also be seen in the forging and extrusion processes with decrease in the specimen size (Tiesler 2002).

Surface Roughness

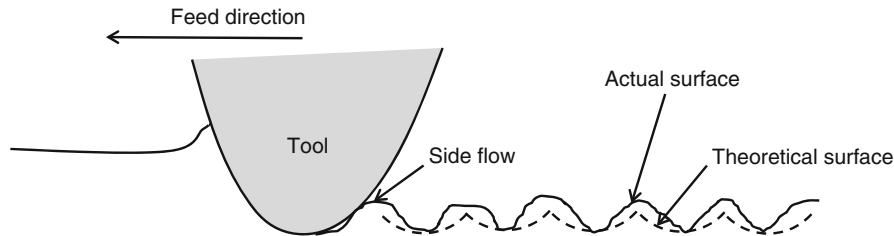
In turning, which is a three-dimensional cutting process, the experimentally measured machined surface roughness first decreases and then increases as the tool feed is decreased (Shaw and Crowell 1965); note that the uncut chip thickness in turning is directly proportional to the tool feed. This size effect in the observed surface roughness is in contrast to theoretical predictions of surface roughness based purely on kinematic and geometric considerations that do not indicate this size effect. Surface roughness in general is affected by machine characteristics (stiffness, vibration, tool clearances, feed control), tool edge geometry and condition, and work material

Size Effects in Machining Tribology, Table 1 Energy per unit volume in machining

Traditional model	$u = u_{\text{plastic}} + u_{\text{friction}}$
Proposed model	$u = u_{\text{plastic}} + u_{\text{friction}} + u_{\text{surface}}$



Size Effects in Machining Tribology, Fig. 7 At small t_o , crack-free surfaces can be obtained in machining a brittle material



Size Effects in Machining Tribology, Fig. 8 Theoretical and actual surfaces in turning; the cutting velocity direction is perpendicular to the page while the feed direction is to the left. The plastic side flow is indicated similar to an indentation pile-up

deformation. The effect of work material deformation on the surface roughness is an important factor. Liu and Melkote (2006b) have shown that the commonly observed discrepancy between the theoretical and measured surface roughness in micro-turning is mainly due to surface roughening caused by plastic side flow during plastic deformation of the material (Fig. 8); the side flow is similar to the pile-up seen in indentation hardness measurements. They also conclude that the typical tool edge quality does not influence the overall surface roughness significantly. Childs et al. (2008) observe that, when machine characteristics are not the limitation, the tool edge radius plays an important role in surface roughness below a certain feed rate; they report a minimum surface roughness achievable in aluminum alloys that is proportional to the edge radius.

Summary

Efforts to understand the different size effect phenomena observed in machining as the uncut chip thickness (or feed) is reduced have yielded better insights into the

science of machining processes. As discussed in this chapter, a number of factors can contribute to the size effect commonly observed in the specific cutting energy. They include factors related to cutting tool geometry (edge radius), sub-surface plastic flow, material strength, and ductile fracture during material separation. Other reported size effects in machining include those observed in the mean coefficient of tool-chip friction and the machined surface roughness. Improved understanding of the size effects in machining has enabled the implementation of more energy efficient machining processes, especially when machining at smaller length scales, e.g., micro-mechanical machining. Furthermore, some of these phenomena can be beneficially utilized in conventional machining practice. For instance, by adjusting the tool edge radius and the uncut chip thickness chip formation is possible in machining of brittle materials leading to good surface finish and surface integrity. Also, understanding and incorporating the effects of material heterogeneity can lead to improved machinability of advanced materials at smaller scales.

Cross-References

- [Brittle-Ductile Transition](#)
- [Crack Growth in Brittle and Ductile Solids](#)
- [Damage Accumulation](#)
- [Friction Modeling for Machining](#)
- [Sliding Wear](#)

References

- M.E. Abdelmoneim, R.F. Scrutton, Sub-surface damage and edge sharpness in finish machining. *Wear* **27**, 35–46 (1974)
- A.G. Atkins, Modelling metal cutting using modern ductile fracture mechanics: quantitative explanations for some longstanding problems. *Int. J. Mech. Sci.* **45**(2), 373–396 (2003)
- W.R. Backer, E.R. Marshall et al., The size effect in metal cutting. *Trans. ASME* **74**, 61–72 (1952)
- P.N. Blake, R.O. Scattergood, Ductile-regime machining of germanium and silicon. *J. Am. Ceram. Soc.* **73**(141), 949–957 (1990)
- T.H.C. Childs, K. Sekiya, R. Tezuka, Y. Yamane, D. Dornfeld, D.-E. Lee, S. Min, P.K. Wright, Surface finishes from turning and facing with round nosed tools. *CIRP Ann. Manuf. Technol.* **57**, 89–92 (2008)
- P.H. Cohen, J.T. Black, J.G. Horne, A.A. Shih, Orthogonal machining of single crystals. *9th North American Manufacturing Research Conference Proceedings*, SME, Dearborn, Michigan USA, 1981, pp. 388–396
- D. Dinesh, S. Swaminathan, S. Chandrasekhar, T.N. Farris, An intrinsic size-effect in machining due to the strain gradient, in *Proceedings of the ASME IMECE*, New York, 11–16 Nov 2001, pp. 197–204
- Y. Furukawa, N. Moronuki, Effect of material properties on ultra precise cutting processes. *Ann. CIRP* **37**(1), 113–116 (1988)
- S.S. Joshi, S.N. Melkote, An explanation for the size-effect in machining using strain gradient plasticity. *Trans. ASME J. Manuf. Sci. Eng.* **126**(4), 679–684 (2004)
- R. Komanduri, Some aspects of machining with negative rake tools simulating grinding. *Int. J. Mach. Tool Des. Res.* **11**(3), 223–233 (1971)
- E.M. Kopalinsky, P.L.B. Oxley, Size effects in metal removal process, in *Third Conference on the Mechanical Properties of Materials at High Rates of Strain*, Oxford, 1984, pp. 389–396
- J. Larsen-Basse, P.L.B. Oxley, Effect of strain-rate sensitivity on scale phenomenon in chip formation, in *Proceedings 13th International Machine Tool Design & Research Conference*, University of Birmingham, Birmingham, 1973, pp. 209–216.
- B.L. Lawson, N. Kota, O.B. Ozdoganlar, Effects of crystallography anisotropy on orthogonal micromachining of single-crystal aluminum. *J. Manuf. Sci. Eng. Trans. ASME* **130**(3), 03116-1–11 (2008)
- W.B. Lee, S. To, C.F. Cheung, Effect of crystallographic orientation in diamond turning of copper single crystals. *Scr. Mater.* **42**(10), 937–945 (2000)
- K. Liu, S.N. Melkote, Material strengthening mechanisms and their contribution to size effect in micro-cutting. *Trans. ASME J. Manuf. Sci. Eng.* **128**(3), 730–738 (2006a)
- K. Liu, S.N. Melkote, Effect of plastic side flow on surface roughness in micro-turning process. *Int. J. Mach. Tools Manuf.* **46**, 1778–1785 (2006b)
- D.A. Lucca, R.L. Rhorer, R. Komanduri, Energy dissipation in the ultraprecision machining of copper. *Ann. CIRP* **40**(1), 69–72 (1991)
- K. Nakayama, K. Tamura, Size effect in metal-cutting force. *Trans. ASME J. Eng. Ind.* **90**, 119–126 (1968)
- C.K. Ng, S.N. Melkote, M. Rahman, A. Senthil Kumar, Experimental study of micro- and nano-scale cutting of aluminum 7075-T6. *Int. J. Mach. Tools Manuf.* **46**(9), 929–936 (2006)
- M.C. Shaw, A quantized theory of strain hardening as applied to cutting of metals. *J. Appl. Phys.* **21**, 599–606 (1950)
- M.C. Shaw, J.A. Crowell, Finish machining. *Ann. CIRP* **13**, 5–22 (1965)
- S. Subbiah, S.N. Melkote, Evidence of ductile tearing ahead of the cutting tool and modeling the energy consumed in material separation in micro-cutting. *ASME J. Eng. Mater. Technol.* **129**(2), 321–331 (2007)
- N.A. Tiesler, Microforming – size effect in friction and their influence on extrusion process. *Wire* **52**, 34–38 (2002)
- S. Venkatachalam, X. Li, S.Y. Liang, Predictive modeling of transition undeformed chip thickness in ductile-regime micro-machining of single crystal brittle materials. *J. Mater. Process. Technol.* **209**(7), 3306–3319 (2009)
- M.P. Vogler, R.E. DeVor, S.G. Kapoor, Microstructure-level force prediction model for micro-milling of multi-phase materials. *J. Manuf. Sci. Eng. Trans. ASME* **125**(2), 202–209 (2003)

Skew Axis Gearing with Tapered Pinion

- [Spiroid® and Helicon® Gearing](#)

Skew Axis Gearing Without Tapered Pinion

- [Spiroid® and Helicon® Gearing](#)

Skid Resistance

- [Tire Friction \(Design, Tire–Road Interactions\)](#)

Skin Abrasion

- [Friction, Human Body: Skin](#)

Skin Tribology

- [Friction, Human Body: Skin](#)

Sleeve Bearing Materials

► Fluid Film Bearing Materials

Slide/Roll Ratio

► Gear Sliding

Slider and Media Characterization

THOMAS R. PITCHFORD

Seagate Technology, Bloomington, MN, USA

Definition

The head/disc interface has a nominal physical spacing between the transducer and media on the order of 10 nm. This spacing requirement places significant constraints on the design of heads, media, and coatings. Associated with this will be the need to characterize the head/disc interface. A survey of measurements important in evaluations of the head/disc interface tribology is presented. The survey includes discussion of characterization of surfaces, parameters relevant to head-disc spacing, and measurement of head-disc interactions.

Introduction: Scientific Fundamentals

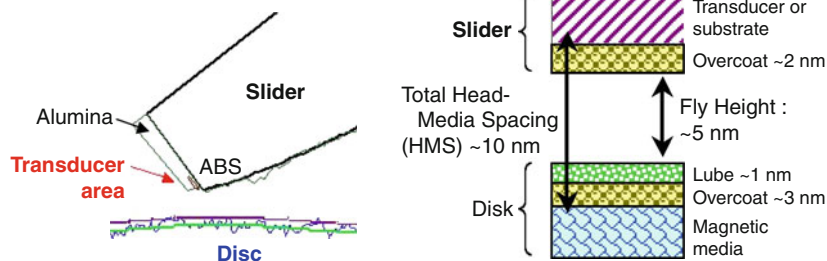
Magnetic disc drives have achieved steady increases in areal density of data storage year after year. One factor in this is reduced spacing between the read/write transducer and recording media. Continued developments in

the tribological design of disc drives have maintained the reliability of the head-disc interface despite decreased spacing. Along with the development of any new designs, there is the need to characterize or measure the system. Tribology measurements of the head-disc interface characterize the mechanical condition of the interface, aiding the development of new designs.

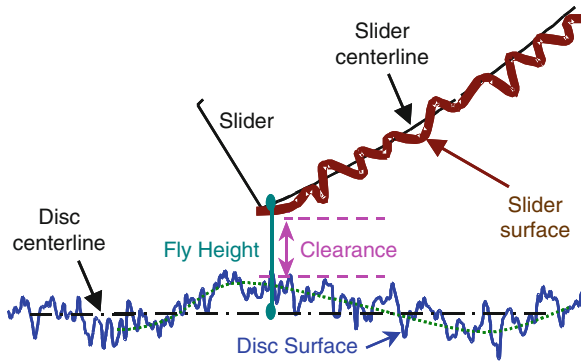
The three main areas of head-disc tribology measurements are characterization of surfaces, head-disc spacing, and head-disc interaction. The design of surfaces will be an important factor in achieving a tribologically sound head-disc interface at ultra low head-media spacing. Spacing measurements are needed to assess how much mechanical margin exists before head-disc contact occurs. Once contact occurs, measurements of head-disc interaction are important in characterizing the degree of stress occurring at the interface. The interface will need to withstand contact that may occur during start/stop and normal operations without undue wear. These factors will be more crucial at ultra low spacing.

As the capacity and performance of disc drives has improved, the mechanical interface has evolved. The spacing between the head and the media has steadily decreased to achieve the rapid improvements in areal density. [Figure 1](#) is a diagram of head-media spacing (HMS). The HMS is comprised of the fly height and coatings (head and disc overcoats, disc lube). Analysis of long-term trends indicates that the HMS between the head and media follows $\sim \text{bit length}/2$ (Marchon and Olson 2009).

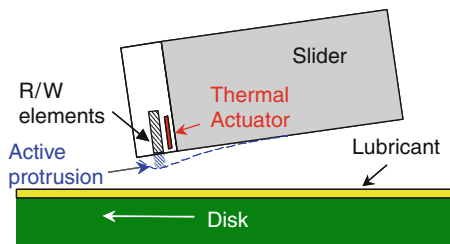
In the diagram of [Fig. 1b](#), the surfaces are assumed to be perfectly consistent and smooth – they do not include factors such as surface topography and variability of the fly height. These would also be on the order of 1 nm. The fly height denotes the spacing between the centerline surfaces of the head and disc. The clearance denotes the spacing between the close points of the surfaces. The distinction between fly height and clearance is depicted in [Fig. 2](#).



Slider and Media Characterization, Fig. 1 Components of head-disc spacing: (a) Schematic of trailing end of head as it flies over disc. (b) Idealized HMS stack-up



Slider and Media Characterization, Fig. 2 Parameters related to head-disc spacing



Slider and Media Characterization, Fig. 3 Active fly height control (Diagram courtesy Bo Liu)

As head-disc spacing has decreased, the means to achieve this has changed. In older drive designs the air bearing design and passive topography of the slider determined the fly height of the transducer. In newer designs the fly height is actively controlled by a signal that changes the shape of the slider. The most common method for this is to embed a heater in the slider that will cause the transducer area to protrude closer to the disc, as is shown in Fig. 3. Developments such as active fly height control have allowed the clearance to decline at a faster rate than other portions of HMS (Marchon and Olson 2009). Understanding the behavior of the slider as the control signal is applied becomes very important as this is required to control the location of the close point and determine the correct signal to achieve the target fly height or clearance at the transducer.

The diagrams of the head-disc interface assume that some degree of clearance will be maintained between the head and disc and that protective overcoats will be employed. While other means of accommodating the magnetic spacing may be developed, the configuration in Figs. 1 and 2 will be the basis for this survey. Depending on the

quality of the head and media, the interface may need to be designed to withstand intermittent or continuous contact.

The following sections present a survey of measurements important in evaluations of the head-disc interface. The sections discuss measurements for three areas of head-disc tribology: (1) [Characterization of Surfaces](#), (2) [Head-Disc Spacing](#), and (3) [Head-Disc Interaction](#). This survey includes many of the characterization methods employed for the INSIC EHDR project (INformation Storage Industry Consortium Extremely High Density Recording). Tribology measurements for 100 Gb/in were surveyed previously (Pitchford 1999). This entry will focus on measurement technologies as areal density approaches 1 Tb/in. There is also a discussion of the effect of external factors on tribology test results.

Key Applications/Characterization Methods

Characterization of Surfaces

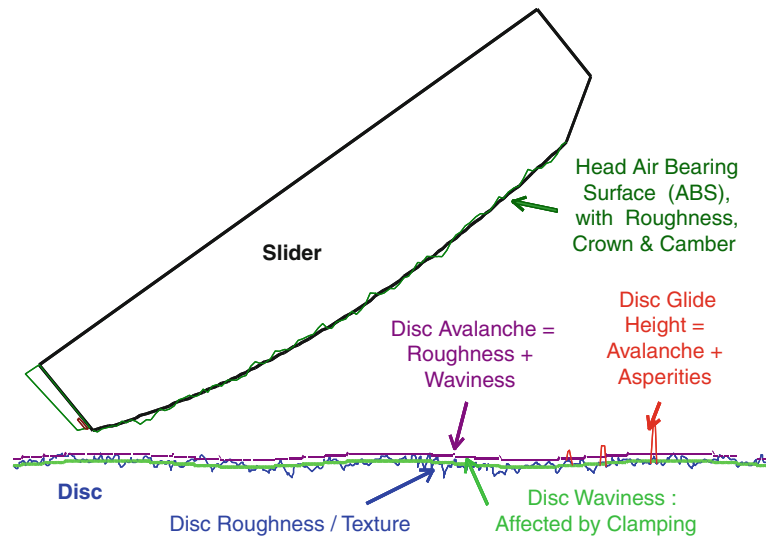
Surface measurements are employed to measure surface topography and characterize the mechanical and chemical properties of materials and surface coatings. For evaluation of head-disc interface performance, surface measurements can be employed to characterize changes that occur during over the course of drive testing or operation.

Surface topography measurements characterize surface flatness, waviness, and roughness. These reflect surface height variations with wavelength ranging from much longer to much shorter than the slider dimension. Figure 4 shows a diagram of the head-disc interface and surface topography parameters.

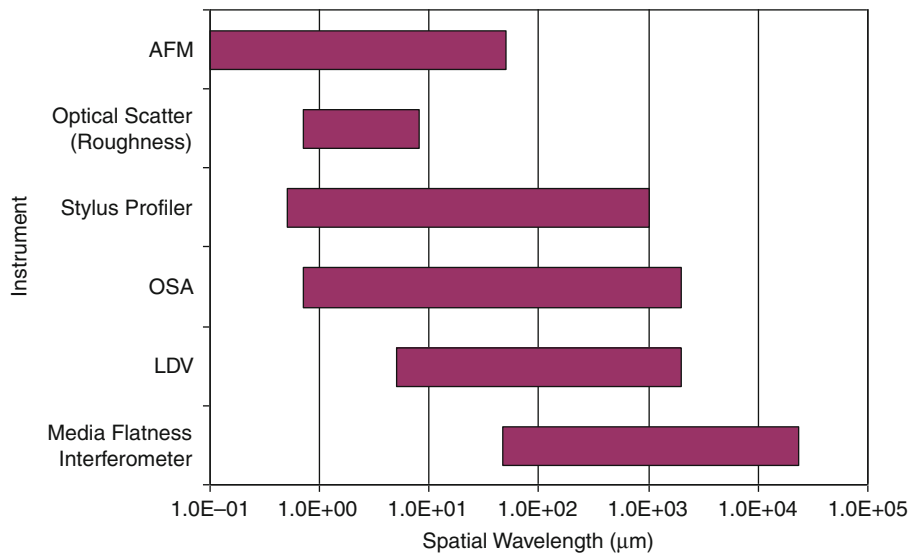
The main methods for characterization of topography are based on optical interferometry (Caber et al. 2000), or scanning probes (AFM, stylus) (Kulkarni et al. 2000). Given the capability for some of these systems to measure atomically flat surfaces, they should be sufficient for characterizing future surfaces. Different instruments are optimized for different ranges of spatial wavelength. The spatial wavelength range of many surface profilers is shown in Fig. 5.

Due to the flatness of surfaces involved in the drive mechanical interface, the contribution of the instrument to the measured profile must be considered. Probe instruments can experience bowing of profiles over the entire length or near edges. The reference surface of interferometers contributes to the measured profile. Care must be taken to minimize or subtract out these contributions.

Once a surface is measured, there are many ways the profile data can be analyzed. Simple parameters such as R_a , R_q , R_p , can be used to predict the clearance between



Slider and Media Characterization, Fig. 4 Surface topography parameters



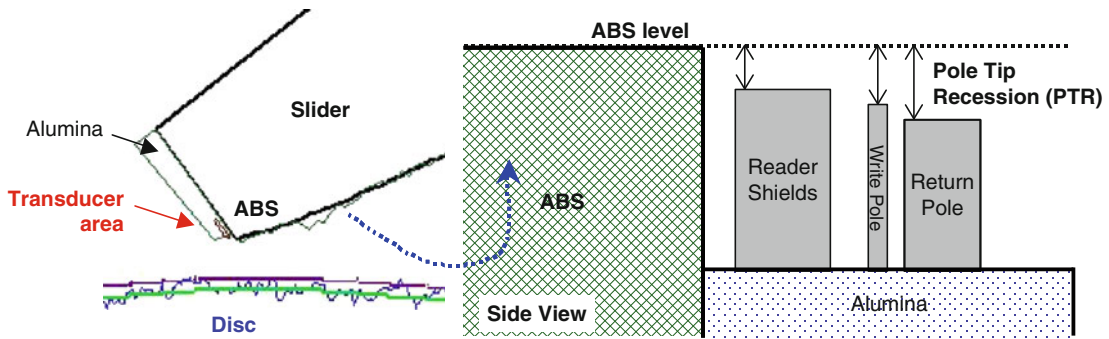
Slider and Media Characterization, Fig. 5 Spatial wavelength ranges for surface profilers

the head and disc. For a given fly height, rougher surfaces will have lower clearance. More detailed parameters such as distribution of peak heights and their radius of curvature are required to predict contact force (Gupta and Boggy 2008, and references therein).

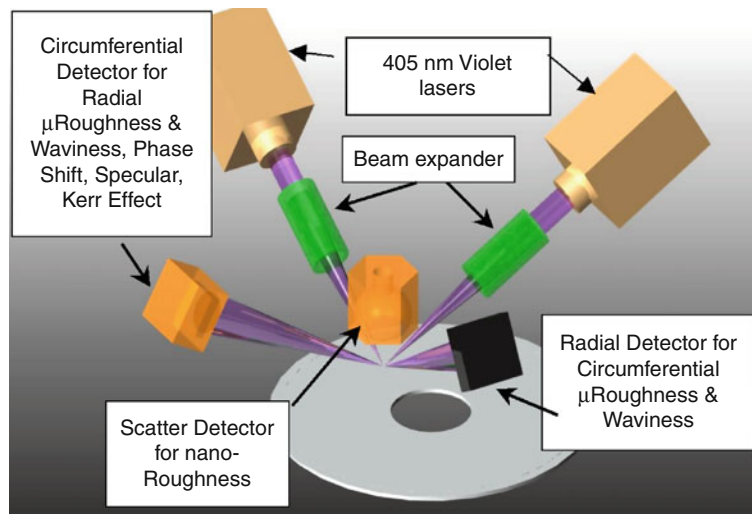
Topography measurements are also employed to characterize isolated surface features. Measurement of asperity height is important in assessing the effect of surface defects or surface damage on head-disc performance. Combining fly height measurement of the slider with measurement of

disc asperity height allows the estimation of interference between the slider and asperities.

As head-disc clearance is reduced and with the use of active control, characterization of the head and disc at the likely point of contact is important. The head will likely contact the disc at the highest point of topography, whether that is with the nominal texture or with an asperity. The point of contact on the head is dependent on the shape of the head, its attitude while flying (pitch and roll), and the local topography of the head in the region of



Slider and Media Characterization, Fig. 6 Slider trailing edge area with transducer details



Slider and Media Characterization, Fig. 7 Diagram of optical surface analyzer (Source: KLA-Tencor/Candella Instruments)

contact. This is true for the nominal profile of the head, and also for the profile in response to the active control and environmental effects such as temperature. Characterization of this local topography requires accurate measurements of all components of the transducer and the surrounding area, as is shown in Fig. 6. Typically the height or recession of the transducer components (write pole, reader shields, etc.) are measured relative to the ABS surface. A parameter called pole tip recession (PTR) could refer to the recession of one of the components or to a weighted average of multiple components.

For accurate assessment of the close point, profilometry methods must be accurate over the ABS, alumina, and transducer areas of the slider. The transducer area has metallic and dielectric components. The ABS portion of the slider could nominally be a single material but is typically a composite of alumina (Al_2O_3) and

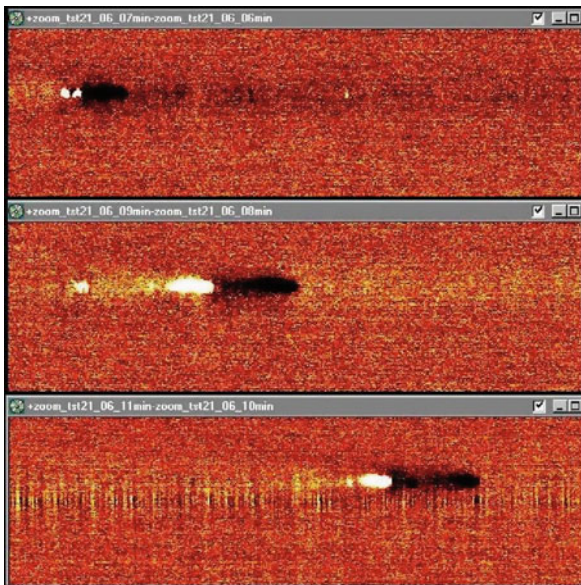
titanium-carbide (TiC). The slider surface is also covered with a protective overcoat. This diversity of materials affects optical measurements. For accurate measurements of parameters such as the pole tip recession (PTR), optical measurements must be corrected based on calibration to AFM or measurements of optical properties of the material components.

During head-disc operation there can be surface wear and damage, lubricant buildup and depletion, lubricant degradation, debris generation, and/or corrosion. A useful tool for characterizing many of these parameters is the optical surface analyzer (OSA). The OSA is an enhanced scanning ellipsometer, as shown in Fig. 7. Polarized light is reflected off of the sample surface. Analysis of the specularly reflected light enables detection of carbon wear and changes in the lubricant. Newer OSA instruments enhance sensitivity by analyzing phase in addition to intensity

(Meeks 2000). The scattered light is used to detect roughness changes, surface debris, dings or scratches. The instrument is sensitive to small changes: 0.1 nm in the case of carbon or lubricant thickness changes. The system can perform quick, high-resolution measurements of entire disc surfaces. The sensitivity of the instrument to changes in coating thickness appear robust relative to differences in deposition processes or precise chemical makeup. While there may be issues with accuracy when multiple factors are affecting the surface simultaneously, the system is still useful for determining locations that require further surface analysis by other methods.

Initially OSAs were employed to check the condition of the disc before and after testing. There is increasing interest in monitoring the condition during testing. This has lead to the development of in situ OSA systems that can perform measurements during head-disc testing. An example of such data showing head-lube interaction is shown in Fig. 8.

A summary of the characteristics of current surface measurements is given in Table 1.



Slider and Media Characterization, Fig. 8 Mapping of changes in lube topography due to head-disc interaction. The OSA images depict the movement of a large lube droplet along a disc track as a head is flown over it. The disc was prepared with greater than normal lube thickness 2 nm. The images encompass a 5-min period for scans over a $350\ \mu\text{m} \times 27^\circ$ area. The *dark areas* denote greater thickness; the lube droplet was up to 40 nm height (Moseley and Bogoy 2009)

Head-Disc Spacing

Head-disc spacing measurements are important in assessing conformance of designs to targets and determining the clearance of the head above the disc. Measurements of average spacing and dynamic spacing variation can be performed. Head-disc spacing measurements for tribology comprise two types: fly height and clearance, as was shown in Fig. 2. Fly height is typically used to characterize the basic head spacing characteristics apart from the disc. It does not include effects due to the topography of the head or disc. Clearance, on the other hand, is a measure of the spacing decrease required to cause head-disc contact.

Fly height is measured by flying the head over a glass disc and employing interferometric techniques. Two common techniques employ either multiple wavelengths of non polarized light or measurement of intensity and phase of polarized light, as is shown in Fig. 9 (Sappey et al. 2006). In either case, a model of the reflected light characteristics versus fly height is used. The intensity and phase of the reflected light can then be related to the fly height. Multiple wavelengths or intensity and phase are employed to prevent ambiguous data.

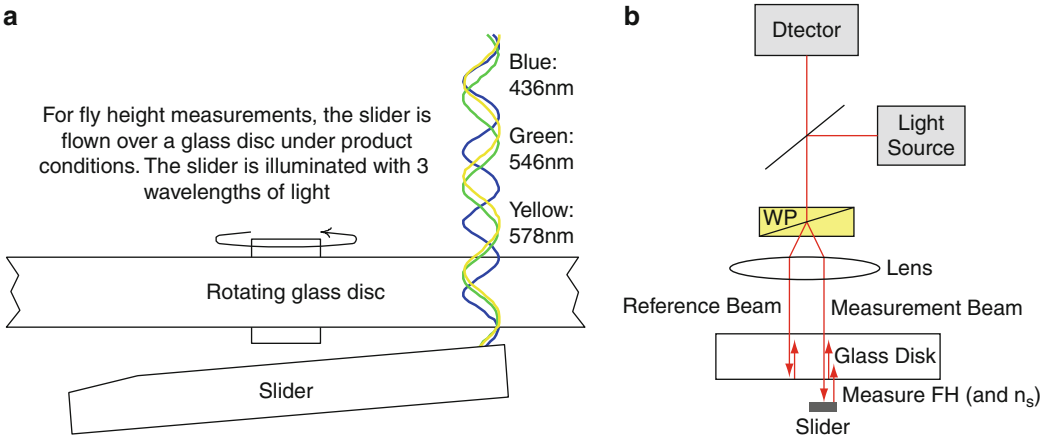
For measurements at low fly height, the signal-to-noise performance of fly height testers may be an issue. Simple optical theory would imply that when the fly height is zero, the intensity is at a local minimum. This would give low measurement sensitivity. However sliders have complex index of refraction $\tilde{n} = n + ik$, where $k \neq 0$. This leads to a shift in the intensity versus fly height curve. This increases the slope of the intensity versus fly height curve near zero fly height. Testers employing phase detection are also immune to SNR issues near zero fly height. Dynamic fly height can be measured on the fly height testers just discussed. It can also be measured with laser Doppler vibrometers. This is discussed further in the next section.

Knowing the values of n and k is required to obtain the correct absolute value of fly height. Originally this required separate ellipsometer measurements. Improvements in testers allowed in situ accounting for optical properties. Another source of error has been the need to perform a retract of the head off of the disc in order to calibrate minimum and maximum intensity values. Negative pressure characteristics of the air bearing can lead to unpredictable dynamics during this retract, making it difficult to capture the full intensity curve. Newer testers do not require this retract calibration.

Optical fly height measurements can be used to calibrate measurements of head-disc clearance, where the onset of contact at any location on the slider is being detected.

Slider and Media Characterization, Table 1 Surface measurements

Type	Measurement	Magnitude/ data	Comment
Topography	Roughness	Ra <0.5 nm	Surface variation over ~ transducer structure
	Waviness	1 nm	Surface variation over ≈ slider air bearing footprint
	Flatness	100 nm	Surface variation over entire disc. For waviness and flatness, contribution of reference surface or profile bowing must be accounted for or reduced
	Asperities	3 nm	Requires ~ 1 μm lateral resolution
	Wear	0.1 nm	Good performance with optical surface analyzer for uniform wear



Slider and Media Characterization, Fig. 9 Comparison of fly height measurement methods: (a) polychromatic intensity based, (b) monochromatic phase based

In this case, knowledge of the fly height of the lowest point on the slider is required, whether it be on the air bearing surface (ABS) or transducer area. The ability to locate and measure the lowest flying point on the slider will be more critical at lower fly heights, especially with active control. Currently, on standard fly height testers, it is only possible to measure on the ABS surface. Due to this, fly height measurements of the ABS, and topography measurements of the transducer area must be combined to estimate the fly height at the likely close point. The read-back signal has been employed to estimate the fly height at the reader element while flying over a real disc (Xu et al. 2006).

In measurements of head-disc clearance, the spacing of the head above the peaks of the disc topography is measured. The clearance is less than the fly height, as Fig. 2 shows. Clearance is typically measured by changing the fly conditions of the head in such a way as to reduce its fly height and induce contact with the disc. The conditions most often changed are disc rotational speed (RPM) or

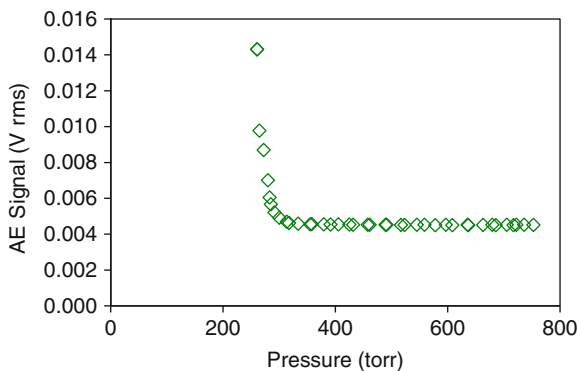
atmospheric pressure. The RPM or pressure at which contact occurs is measured. A curve of head fly height versus RPM or pressure is required and is typically obtained by measurement. This enables the conversion of the RPM or pressure corresponding to contact to a fly height value. Other conditions that may be changed are temperature or humidity. With active fly height control, the control signal can also be used to change clearance of the transducer area.

A sensor is required to detect the onset of head-contact. For example, an acoustic emission (AE) sensor is used to detect head vibrations caused by contact. An example of a measurement of AE signal versus pressure is shown in Fig. 10. For high pressures the AE signal is at its baseline value. The data indicates contact at around 300 torr. Below this pressure there is a sudden increase in the AE signal. This increase is referred to as an avalanche and this measurement method is known as an altitude avalanche. Similar data could be obtained by varying

RPM. The fly height corresponding to the avalanche is often referred to as the take-off-height (TOH).

There are many sensors that will give indication of head-disc contact. These may sense anomalies in the motion of the slider in the off-track, down-track, and/or vertical direction. External sensors include AE, PZT, friction, and LDV. Internal sensors employ the sliders own electrical signal to sense changes in clearance or the onset of contact – signals include amplitude, PW50, harmonic ratio, off-track, and jitter.

The avalanche measurement is typically used to characterize the margin for clearance between the head and the intrinsic texture of the disc. In this case the normal product head is used in the test. The normal avalanche test setup is typically not sensitive enough to detect isolated asperities with adequate signal to noise. Detection of contact with disc asperities is accomplished with so-called glide height testing. For glide testing a special head with PZT material is attached to the top of the slider. The PZT material generates a signal in response to asperity contact. The noise in glide height measurements will probably be comparable to avalanche measurements.



Slider and Media Characterization, Fig. 10 Altitude avalanche measurement

A summary of head-disc spacing measurements is given in Table 2. The table lists the measurement types and measurements discussed. The third column, magnitude/data, gives the estimated mean value expected for the parameter. The comment section summarizes the prospects for the measurement where applicable in terms of accuracy, repeatability, and ability to discriminate changes.

Head-Disc Interaction

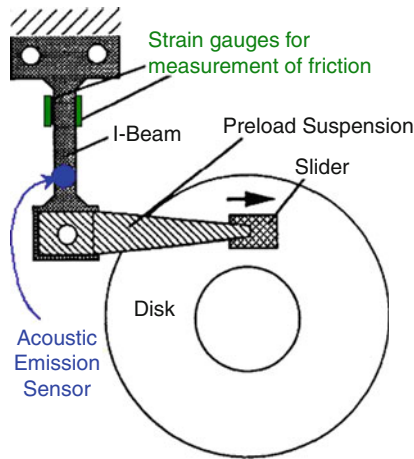
Measurements of head-disc interaction characterize interface dynamics while the head is flying or in contact with the disc. This is important in verifying the stability and durability of the interface during normal operation or in conditions of intermittent to continuous contact. To characterize head-disc interactions, many sensors can be employed – often the same sensors employed in detecting contact measuring clearance. These may sense anomalies in the motion of the slider in the off-track, down-track, and/or vertical direction. External sensors include AE, PZT, friction, and LDV. Internal sensors employ the sliders own electrical signal to sense changes in clearance or the onset of contact – signals include amplitude, off-track, and jitter. Some of the sensors such as LDV and electrical amplitude are well suited for measurements under flying or contact conditions. Other sensors such as AE and off-track are better suited for just contact.

A typical setup for measurement of head-disc interaction is shown in Fig. 11. The head is mounted on a special arm. Strain gauges are mounted on the arm to detect deflection due to friction between the head and the disc. A good understanding of tester mechanics is required for meaningful friction measurements (Li and Menon 1994). Measurement of friction can be useful in characterizing head-disc interaction even for designs that do not experience contact-start-stops. The setup can also include an acoustic emission (AE) sensor, as shown.

As head-disc spacing decreases, the magnitude of signals (friction, AE) arising from head-disc contacts has not

Slider and Media Characterization, Table 2 Measurement of head-disc spacing

Measurement type	Measurement	Magnitude/data	Comment
Fly height	Head fly height	5–10 nm	Estimated measurement uncertainty 0.2 nm (1σ), accuracy linearity errors 0.5 nm
Clearance	Head/disc avalanche or takeoff height (TOH)	1–3 nm	Overall variability of measurement 0.5 nm (1σ) including contribution of head
	Disc glide height	3–5 nm	Results comparable to those for head-disc avalanche



Slider and Media Characterization, Fig. 11 Typical measurement setup for measurement of breakaway or dynamic friction (After Li and Menon 1994)

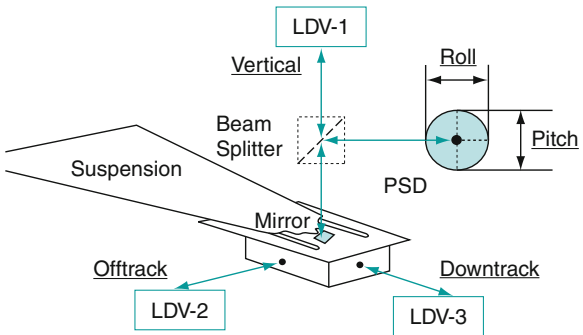
changed significantly. But as surfaces become smoother, the magnitude of head-disc interaction has tended to increase (Thornton and Bogoy 2003). With the decreased head-disc spacing, there will be less margin between flying and contact, possibly leading to higher incidence of head-disc contacts. There is a need to understand and minimize the forces associated with contact. Characterization of contact forces requires high bandwidth measurements of slider vibrations due to contact. Progress has been made in methods for calibration and measurement of head-disc contact forces (Ganapathi et al. 1995).

The slider geometry has been moving from so-called nano to pico to femto sliders. As slider size has decreased, the bandwidth required for measurements has increased. The main vibrational modes of various slider form factors are compared in Table 3. Improvements in AE sensors have increased their bandwidth, but the attenuation of the signal through the suspension is still a limitation. This is not as much of an issue with detection of contact with the disc texture.

Besides characterization of interaction through friction or AE measurements, direct measurement of head-disc motion can be performed. The laser Doppler vibrometer (LDV) can be used to sense head and disc motion, as well as the differential motion of the slider relative to the disc. This is important in determining the transfer function characteristics of the interface components such as suspension and air bearing. Multiple LDVs can be combined with position sensitive detection (PSD) to measure three axes of motion plus pitch and roll of slider dynamics during contact. The setup and some data from this system are shown in Figs. 12 and 13, (Hsia and Kiely 2006).

Slider and Media Characterization, Table 3 Slider vibrational modes for IDEMA standard form factors. Values in MHz

Mode	nano	pico	femto
1st: torsion about the long axis	0.71	1.3	1.9
2nd: bending about the short axis	0.92	1.7	2.6
3rd: bending about the long axis	1.5	2.6	3.3



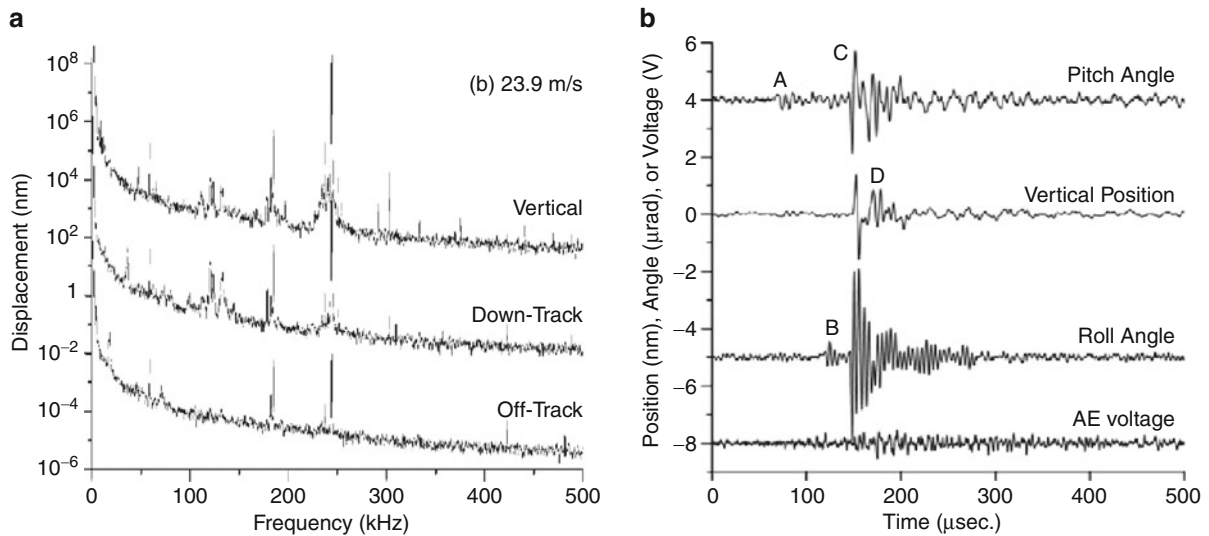
Slider and Media Characterization, Fig. 12 Five channel system for measurement of three axes of motion plus pitch and roll in response to head-disc contact

A summary of measurements of head-disc interaction technology is given in Table 4.

Besides the choice of sensors, the method of data collection and analysis is always important. The most basic method is time based for analysis of temporal changes in the signal. For frequency analysis the data can be converted to a spectrum for analysis of signal components. Components of the interface such as the suspension, slider, air bearing, and disc have characteristic resonant frequencies that can be discerned in a spectrum. For example in Fig. 13a, the peak at ~240 KHz corresponds to the pitch mode of the air bearing. The most sophisticated method is joint-time-frequency analysis, which can analyze changes in the frequency content of a signal versus time or circumferential location (Thornton and Bogoy 2003).

Effect of External Factors on Testing

In the previous sections the characteristics and requirements of many measurement techniques have been discussed. With the low head-disc spacing envisioned in the future, control of external factors will be more crucial in tribology testing. This is especially true for



Slider and Media Characterization, Fig. 13 Data from five-channel sensing system. (a) Power spectrum of the three axes of slider lateral motion while in light contact with media. (b) Slider vertical and angular LDV signal and AE response corresponding to head contact with media defect (Data offset for clarity)

Slider and Media Characterization, Table 4 Measurements of head-disc interaction

Measurement type	Measurement	Magnitude/data	Comment
Contact	Friction	10 mN	Current methods adequate
	Acoustic emission		Sensitivity to 3 + MHz
	Contact force	1 mN	Requires high bandwidth and dynamic range AE. Sensor on slider may be required
Spacing	Dynamics	0.5 nm	Current methods (LDV, read element) adequate

measurement of head-disc spacing and head-disc interactions. If not controlled, external factors can be the main source of measurement variability. This would compromise the meaningfulness of any results.

The distortion of the head-gimbal assembly (HGA) or disc due to mounting on the tester can be significant. The methods and mechanics for mounting of HGAs and disc on testers may not satisfy advancing requirements for control of distortion. Clean room and environmental specifications must keep pace with drive requirements. At the same time external vibrations must be controlled. For fly height testing, there must also be improvement in the finish of glass discs used for fly height testing.

External effects that cannot be controlled must be characterized. This would entail characterization of all parts and testers before test. Parameters to be characterized would include the mechanical parameters of the HGAs and disc in their tested configuration. Parameters include HGA preload, slider flatness, and disc flatness as clamped on the tester. Parameters that might change during testing must be measured pre- and post-test.

Conclusion

Tribology measurements are crucial in ensuring a reliable head-disc interface. The three main areas of head-disc tribology measurements are characterization of surfaces,

head-disc spacing, and head-disc interaction. While the precise design of future interfaces is still being defined, it is evident that the design of head-disc spacing, materials, surface texture, protective coatings, and lubricants will be significantly more constrained than current designs. In general, current methods for measurements will continue to be satisfactory for measuring relative changes due to changing external conditions (such as RPM or pressure), or changes that occur during testing (such as wear or lube thickness). Increased use of on-slider sensors may be required. The surface topography measurements will be more influenced by the contribution of the reference surface or bowing. Improved control or measurement of external factors that affect tribology measurements will be more crucial for meaningful measurements.

Cross-References

- [Disk Roughness and Defect Monitoring](#)
- [Surface Texture for Magnetic Recording](#)

References

- P. Caber, G. Artur, A. Olszak, C. Ragan, Present and future interference microscope systems for magnetic head metrology. In *Proceedings of SPIE*, vol. 4099 (2000) p.166
- S.K. Ganapathi, M. Donovan, Y.-T. Hsia, Contact Force Measurements at the Head/Disk Interface for Contact Recording Heads in Magnetic Recording. In *Proceedings of SPIE*, vol. 2604, 23–24 Oct 1995
- V. Gupta, D. Bogy, Optimal slider-disk surface topography for head-disk interface stability in hard disk drives. *IEEE Trans. Magn.* **44**(1), 138 (2008)
- Y.-T. Hsia, J. Kiely A novel metrology technique to capture the flying dynamics of air bearing slider with and without induced contact. Asia-Pacific Magnetic Recording Conference, Nov 2006, p. 1
- A. Kulkarni, S. Chilamakuri, B.K. Gupta, A. Menon, Pole tip recession (PTR) measurements with high accuracy, precision, and throughput. *IEEE Trans. Magn.* **36**(5), 2736 (2000)
- Y. Li, A. Menon, Theoretical analysis of breakaway friction measurement. *J. Tribol. Trans. ASME* **116**(2), 280–286 (1994)
- B. Marchon, T. Olson, Magnetic spacing trends: from LMR to PMR and beyond. *IEEE Trans. Magn.* **45**(10), 3608–3611 (2009)
- S. Meeks, Combined ellipsometer, reflectometer, scatterometer, and Kerr effect microscope for thin film disk characterization. In *Proceedings of SPIE*, vol. 3966 (2000) p. 385
- S. Moseley, D. Bogy, Experimental evidence of lubricant droplet transfer from slider to disk. *IEEE Trans. Magn.* **45**(2), 867 (2009)
- T. Pitchford, Head/disk interface tribology measurements for 100Gb/in, interface tribology towards 100 Gb/in. In *ASME symposium*, 10 Oct 1999, p. 83
- R. Sappey, T. Carr, M. Loera, C. Lee, Phase-sensing interferometry for flying-height metrology. Digest of the Asia-Pacific Magnetic Recording Conference (2006), p. 1
- B. Thornton, D. Bogy, Nonlinear aspects of air-bearing modeling and dynamic spacing modulation in sub-5-nm air bearings for hard disk drives. *IEEE Trans. Magn.* **39**(2), 722 (2003)
- J. Xu, J. Kiely, Y.-T. Hsia, F. Talke, Head-medium spacing measurement using the read-back signal. *IEEE Trans. Magn.* **42**(10), 2486 (2006)

Slideway Lubricant

- [Slideway Lubricants](#)

Slideway Lubricants

IAN MACPHERSON, JOHN TURTLE, JOHN M. TAYLOR
Afton Chemical Corporation, Richmond, VA, USA

Synonyms

[Slideway lubricant](#); [Way lube](#); [Way lubricant](#); [Way oil](#)

Definition

The effective lubrication of equipment used in slideway operations is summarized. This includes discussion of the specific lubricant chemistry involved, together with some of the performance tests used to qualify a lubricant for this application. While the variety of design and composition of slideways is complex, there are some common lubrication performance needs that are essential for efficient operation.

Scientific Fundamentals

Slideways or ways are used as the medium on which to slide heavy equipment. This type of system is used when the equipment to be transported is either very heavy or when the movement requires high precision. Good examples of slideways are seen within the industrial manufacturing sector, where extremely heavy machine tools are positioned very precisely to the work pieces. This positioning may involve horizontal motion, vertical motion, or both. During use, slideways and slideway lubricants are required to cope with several performance challenges, including frictional properties, adherence properties, coping with contamination, and corrosion resistance, which are discussed below.

While the following general properties are important for effective slideway lubrication, selection of the specific slideway lubricant must be performed in accordance with the equipment builder's recommendations or requirements ([Cincinnati Machine](#)).

Viscosity

The ISO viscosity grade nomenclature ([ISO 1992](#)) is usually used to define the viscometric requirements of slideway lubricants. This system defines the kinematic viscosity

of the lubricant at 40 °C. The most common viscosity grade is ISO 220, which means that at 40 °C the kinematic viscosity of the oil will be at $220 \text{ mm}^2/\text{s} \pm 10 \%$. In addition, ISO 32 and ISO 68 viscosity grades are also common. Low-temperature viscosity requirements are usually not a consideration, other than to ensure that the product remains liquid during shipping and storage and can be satisfactorily pumped by the slideway lubrication system. For this reason, slideway lubricants may contain up to around 0.1 % of a polymeric chemical such as an *alkyl*-polymethacrylate, which is designed to prevent any wax from crystallizing at lower temperatures. These polymeric materials are referred to as pour point depressants. The viscometric properties of the lubricant will change with temperature. A convenient measure of these changing properties is expressed in terms of viscosity index (VI). The viscosity index of a typical slideway lubricant is usually around 95–100. It is not common to add any polymeric viscosity index improvers to slideway lubricants. These types of polymers tend to be used only when extremes of high and low temperature are experienced.

The base oils selected for slideway applications vary considerably. These may consist of group I, II, III, and/or IV base oils. In order to provide the higher viscosity grades, brightstock is frequently employed. Interestingly, the extra lubricity afforded by the brightstock may also improve the frictional characteristics of the slideway lubricants, but on the negative side, may impact oxidation control and therefore the formulator's selection of oxidation inhibitors.

Frictional Characteristics

One of the key features of the slideway lubricant is its frictional properties. Indeed, friction control is critical to the entire function of the slideway. The slideway operates by allowing two parallel, plane surfaces to glide over each other with smooth transition from stationary to sliding together with minimum resistance or shuddering. There are several factors that influence this smooth operation.

Slideway Composition

Slideways are usually cast iron or steel, which may also be coated with a friction reducing material. These materials include other metals such as molybdenum or bronze, ceramic materials, PTFE, or other synthetic organic-based resins. There are many manufacturers of these coating materials. The composition of the mating surfaces during the slideway operation has a major influence on the friction. The composition also dictates the effectiveness of the friction-modifying chemistry used in formulating the slideway lubricant.

Contamination

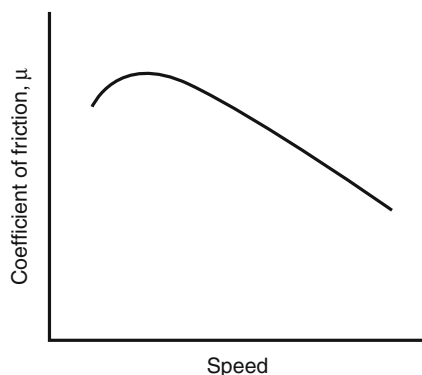
Metalworking machine tools are a common application for slideways and are among the most difficult applications for slideway lubricants. Metalworking fluids, both neat oils and water-based coolants, often contaminate the slideway and the slideway lubricant. This challenges the frictional performance of the slideway as these contaminant fluids will typically have poor frictional characteristics, and may even effectively wash off the lubricant from the slideway. Efficient coolant separation properties of the slideway lubricant are therefore a critical lubricant performance feature, which is discussed in detail below.

Friction Modifiers

Friction modifier is a general term for lubricant additives that are an ingredient of the slideway lubricant specifically added to enhance the frictional properties of the fluid. They are typically surface-active materials with polar heads and hydrocarbon tails. They operate by forming a film on the surface of the slideway, which helps to keep the mating surfaces apart. It should be noted that the slideway lubricant itself comprises greater than 90 % base oil, which by hydrostatic process will form a protective film and help to keep the surfaces apart. Under moderately loaded conditions, however, this protective film dissipates. The strong polar attraction of friction modifiers for the metal slideway surfaces provides greater oil film stability than base oil alone. The selection of friction modifier chemistry is dependent on the specific application and on the friction material that comprises the slideway surface.

Friction Profile

When the relative speed of the slideway surfaces is fairly high, the system is said to be under dynamic friction conditions. As the relative movement slows to a stop, the system is said to be under static friction conditions. The static and dynamic friction coefficients are rarely equal with the static value, usually exceeding the dynamic value unless an effective friction modifier is added to the lubricant. The friction properties of the lubricant are optimized so that this ratio of static to dynamic friction is carefully controlled (Fig. 1). If the coefficient of friction under static conditions is substantially higher than under dynamic conditions, then the system will likely shudder as it comes to a stop. This shuddering or "stick-slip" phenomenon reduces the ability of the controlling system to accurately position the load on the slideway. A high static coefficient of friction also increases the resistance of the slideway to start moving. This can cause sudden surges of



Slideway Lubricants, Fig. 1 Graph showing atypical variation of friction at different speeds of a slideway lubricant (steel on steel)

speed at the start, which again impacts the efficiency and accuracy of the sliding operation.

There are three tests that have found favor among lubricant suppliers and users for the evaluation of the friction and stick-slip properties of slideway lubricants. These are the Cincinnati Milacron (C-M) stick-slip test, the Schmidt inclined tribotester, and the TU Darmstadt slideway friction test. Of these, the C-M stick-slip test is globally most recognized by OEMs and end users.

Adherence or Tackiness

Some slideways are vertical rather than horizontal so there is a practical challenge to keeping the lubricant in the place where it is needed. In addition, the contamination of the slideway lubricant with water-based coolants and cutting oils tends to effectively wash the slideway lubricant from critical sites. The use of ISO 220 grade lubricants on vertical slideways assists in keeping the lubricant in place. *Tackiness*, which is a term used to describe the adherence property of the lubricant to the various mechanical parts, is needed to help the lubricant stay in place as much as possible. Tackiness lubricant additives tend to comprise very-long-chain oil-soluble polymers.

Tackiness Tests

There have been several attempts to measure the degree of tackiness of lubricants in laboratory tests. None of the test methods have been universally accepted. The degree of tackiness is largely dictated by the type and amount of tackiness additive used in the lubricant, and the specific needs of the application. This is often determined by field experience, but usually ranges from 1 % to 5 % in concentration.

Water and Coolant Separation

There is a tendency for large amounts of water-based cutting oils and coolants to be present in many of the applications where the slideway lubricant is used. It is inevitable that some of this fluid will contaminate the slideway lubricant. In order to ensure efficient operation of the facility, good separation is important. It is also important to the efficient operation of the metalworking fluid that any contaminating slideway fluid separates readily from it. This allows effective lubrication and tool cooling to continue.

These separation characteristics of a slideway lubricant from water-based contaminants are often tested by examining their separation properties from distilled water, measured by the procedure defined by ASTM D 1401 ([ASTM Standards Worldwide Petroleum Products and Lubricants](#)). This test may be performed either at 54 °C or at 82 °C, depending on the viscosity grade examined.

Separation characteristics from the cutting oils themselves are also examined by mixing samples of the slideway lubricant with the cutting fluid under controlled conditions. The degree, appearance, and speed of separation are recorded. Coolant compositions vary considerably depending on the application, manufacturer, concentration, and water hardness, so typically separation from a variety of coolants is evaluated as part of the development process.

Careful choice of the additives and base oils used in a slideway lubricant is critical to ensuring effective separation from coolants and an ability to resist contamination. In some case, additional surfactants may also be added to further boost performance.

Wear and Extreme Pressure Protection

The high loading experienced on slideways requires that the lubricant contain additives that prevent wear or other deformation of the surfaces under conditions where the hydrodynamic or hydrostatic oil film breaks down. This type of lubricant protection is often referred to as boundary lubrication.

Boundary lubrication is provided by certain chemical additives used to formulate the lubricant. These typically comprise chemistry that is rich in sulfur, phosphorus, or zinc, though other elements may also be used. Selection of these chemical performance additives is determined not only by their ability to prevent wear and scuffing under boundary conditions, but also by their ability to filter and afford enough thermal stability to give the lubricant longevity. Boundary lubricant additives typically function by chemically interacting with the surface at the critical sites. Often heat is required to activate these boundary

additives, and so only at those sites prone to hot spots (typically the sites where wear is likely) will the chemistry be active.

Performance tests used to evaluate the ability of the slideway oil to provide boundary lubrication include the DIN 51354 ([DIN Deutsches Institut für Normung](#)) FZG load stage test, ASTM D 4172 ([ASTM Standards Worldwide Petroleum Products and Lubricants](#)) 4-ball wear test, DIN 51350 ([DIN Deutsches Institut für Normung](#)) 4-ball weld test, ASTM D 2782 ([ASTM Standards Worldwide Petroleum Products and Lubricants](#)) Timken test, and others defined by slideway or gear equipment manufacturers.

In addition, many operations use the slideway lubricant to lubricate associated gears and even some hydraulics.

Corrosion Resistance

Preventing the formation of rust or corrosion that can easily occur in the wet and humid atmosphere of a machine tool slideway is another property afforded by the slideway lubricant. Certain components of the system may be comprised of yellow metal such as brass or bronze. The slideway lubricant is required to be compatible with these hardware components and to protect them from corrosion. This is particularly important where the lubricant contains sulfur, which may be present in the base oil or in the performance additives selected. Indeed, some antiwear or extreme pressure performance additives comprise large amounts of sulfur as their primary active element.

Corrosion tests used to evaluate the protection afforded by the slideway lubricant include the ASTM ([ASTM Standards Worldwide Petroleum Products and Lubricants](#)) D 665 procedure A and B rust tests and the ASTM ([ASTM Standards Worldwide Petroleum Products and Lubricants](#)) D 130 copper corrosion test.

Rust inhibitors generally are surface active chemicals that cover the metal surface, precluding the penetration of water. Great care is taken when formulating slideway lubricants to ensure that the rust inhibitor does not prevent the antiwear (boundary) additives or the friction modifiers from contacting the surface sufficiently to serve their function. This is an example of a situation where the additive components employed must be carefully balanced to ensure effective performance in multiple areas (rust, wear, and friction).

There are generally two classes of yellow metal corrosion inhibitors. One functions similarly to rust inhibitors, forming a protective film on the metal surface, while the other works as a scavenger. A good example of this latter

class is represented by the *alkyl*-thiadiazoles, which consume active sulfur and prevent it from corroding alloys rich in copper such as brass or bronze.

Foaming Tendency

Foam can lead to many problems in industrial applications if not controlled. Some of these problems relate to safety considerations should the foam spill over onto the floor and cause the area to become a slipping hazard. Foam is a poor lubricant so, should it be circulated with the oil through the system, serious problems associated with lack of lubrication may result. Foam may also serve as a good thermal insulator and increase the overall operating temperature of the lubricant. In some equipment design, the presence of foam in the reservoir may fool the operator with respect to the amount of lubricant in the system. This may lead to running with lower lubricant levels, which in turn may increase the operating temperature, leading to potential oxidation issues. In addition, low lubrication levels may reduce circulation time, which leads to – more foam!

Foam inhibitors used in the oil include those based on silicon chemistry and those based on organic polymer-type chemistry. Each type can be very effective at reducing the tendency for slideway lubricants to foam. Standard foam tests include the ASTM D 892 test ([ASTM Standards Worldwide Petroleum Products and Lubricants](#)), which is run under a variety of temperatures and conditions to confirm adequate resistance to foam tendency and stability.

Great care must be taken in adding foam inhibitors to any lubricant. The effective treat rates are often extremely low – sometimes just a few parts per million (ppm). Adequate dispersion of the inhibitor in the oil can be challenging and may require special high-speed mixers. Surprisingly, addition of too much foam inhibitor can actually lead to higher levels of foam in the system.

Key Applications

Slideways oils lubricate the medium on which equipment is conveyed where wheels or rollers are impractical. This type of system is used when the equipment to be transported is either very heavy or when the movement requires high precision. Good examples of slideways are seen within the industrial manufacturing sector, where extremely heavy machine tools are positioned very precisely to the work pieces. This positioning may involve horizontal motion, vertical motion, or both. There are very few lubricant specifications in the industry for slideway lubricants ([Cincinnati Machine](#)). Typically, the machine builder or the specific application needs drive

the performance requirements. Typical performance requirements vary and are described above. In some applications, the slideway lubricant may also be used to lubricate the hydraulics or even gears associated with the equipment or manufacturing facility. In these cases, additional performance parameters may be important for the fluid.

The slideways upon which the slideway oil lubricant performs its function, may be made of a variety of materials, including iron, steel, and a variety of modern synthetic plastic-type coatings. The tendency for lower and controlled friction, especially in the presence of contamination, remains an important technology driver for materials and the lubricant.

Cross-References

- Additive Chemistry Testing Methods
- Cutting Fluids and Their Environmental Impact
- Friction Modifiers

References

- ASTM Standards Worldwide Petroleum Products & Lubricants
Cincinnati Machine, Hebron, KY 41048. Special Manual Lubricants,
Purchased Specifications, Approved Products; Publication Number
10-SP-00038-2, Part number 9703432212C
- DIN Deutsches Institut für Normung. V. DIN Handbook
International Organization for Standards Publication ISO 3448:1992

Sliding Contact Materials

- Brush Materials

Sliding Electrical Contact Wear

KOICHIRO SAWA
Department of System Design Engineering, Keio
University, Kohoku-ku, Yokohama, Japan

Definition

Sliding electrical contact wear is a material loss caused by sliding friction and sometimes by electrical arcing under repeatedly applied stress and effect of electrical current. The wear is generated both in the brush and the commutator/slip ring.

Scientific Fundamentals

Generally, brush wear has the following three types:

- (a) Mechanical wear with no load current
- (b) Mechanical wear with load current
- (c) Electrical wear (by arc discharge and other causes)

Theoretical Aspects of Wear

It is well known that the real contact area is very small compared with the apparent contact area, as shown in Fig. 1 (Bowden and Tabor 1954; Holm 1967).

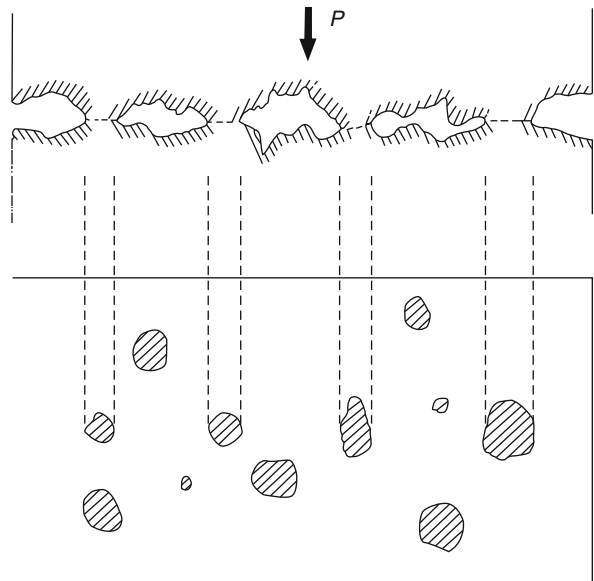
Holm introduced the idea of adhesive wear and proposed an equation for the wear:

$$V = Z \frac{P\ell}{p_m}, \quad (1)$$

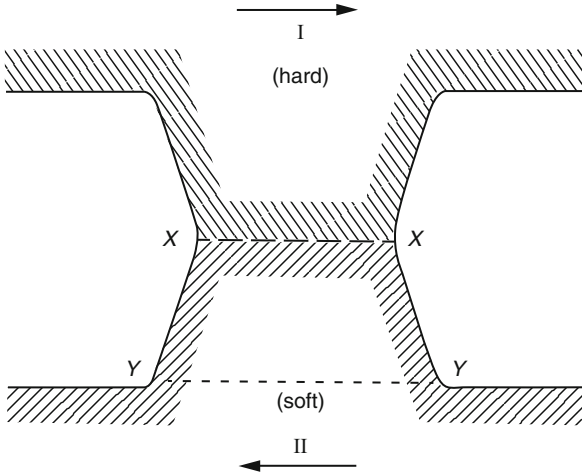
where V : wear volume, P : contact load, ℓ : sliding distance, p_m : hardness, and Z : a constant called wear factor by Holm.

The equation is sometimes called Holm's law of wear, and is regarded as one of the fundamental laws of tribology, together with Coulomb's law of friction.

Rabinowicz and Archard stated for the first time that wear is mechanical rupture of junctions (Rabinowicz 1965). Small real contacts form adhesive junctions, as shown in Fig. 2, and soft projection XXYY is ruptured as a dropping off particle due to the sliding motion. The wear



Sliding Electrical Contact Wear, Fig. 1 Contact surface model



Sliding Electrical Contact Wear, Fig. 2 Adhesive wear model

is proportional to the real contact area and sliding distance as below:

$$V = ZA_R \ell \quad (2)$$

And A_R is obtained as follows:

$$A_R = \frac{P}{p_m} \quad (3)$$

Equation (4) can be derived based on (2) and (3). Equation (4) is the same as Holm's equation of wear

$$V = Z \frac{P \ell}{p_m}, \quad (4)$$

where the constant Z is a probability of wear particles dropping off at the rupture of the junction.

Sasada proposed another wear model called the transfer and growth model (Sasada et al. 1981). According to the model, two small projections make contact on sliding surfaces (Fig. 3a), and internal rupture generally occurs in a soft material (Fig. 3b). A ruptured small particle is attached to a hard material, called a transfer element (Fig. 3c). During slide motion another internal rupture occurs on the previous transfer element (Fig. 3d) and a transfer particle is formed (Fig. 3e). The transfer particle grows larger, as shown in (Fig. 3f), and this larger particle drops off soon thereafter as a wear particle.

In this model it is assumed that there are N real contact spots on an apparent contact area and N spots have the same radius a . When a slider moves by $2a$, all N spots are ruptured and new N spots are generated. Therefore, $\frac{N\ell}{2a}$ spots are ruptured within sliding distance ℓ .

Assuming that a small volume ΔV is removed from a bulk contact with every rupture of a real contact, all

volume V of removed particles within the slide distance ℓ is as follows:

$$V = N \ell \frac{\Delta V}{2a} \quad (5)$$

Further, real contact area (sum of N spots) A_R is

$$A_R = N \pi a^2. \quad (6)$$

Equation (7) is derived from (3), (5), and (6).

$$V = \frac{\Delta V}{2\pi a^3} \frac{P \ell}{p_m} \quad (7)$$

Equation (7) is the same as Holm's equation (1).

In actual wear, the wear rate is higher at the initial stage of sliding motion and decreases to a small value after a certain amount of sliding distance. The initial stage of high wear rate is called "initial wear" and the wear behavior is severe wear where the wear particle is large in size and metallic. On the other hand, the following low wear stage is called "stable wear" and mild wear, where the wear particle is a fine oxide.

Wear Under Conditions of Current and Arc Discharge

At electrical contacts, a current flows at small junctions and the current is constricted at each junction. Thus, each junction is heated by the constriction resistance and the probability of a wear particle dropping off may increase. Therefore, the wear with current would be larger than without current.

Experimentally, mechanical wear with no current is usually one-half or one-third of the wear with current.

Further, the brush wear is dependent on surface conditions of the mating commutator or slip ring, such as roughness, surface film, and so on. Severe wear occurs at very low humidity in the atmosphere (Zaidi et al. 1990).

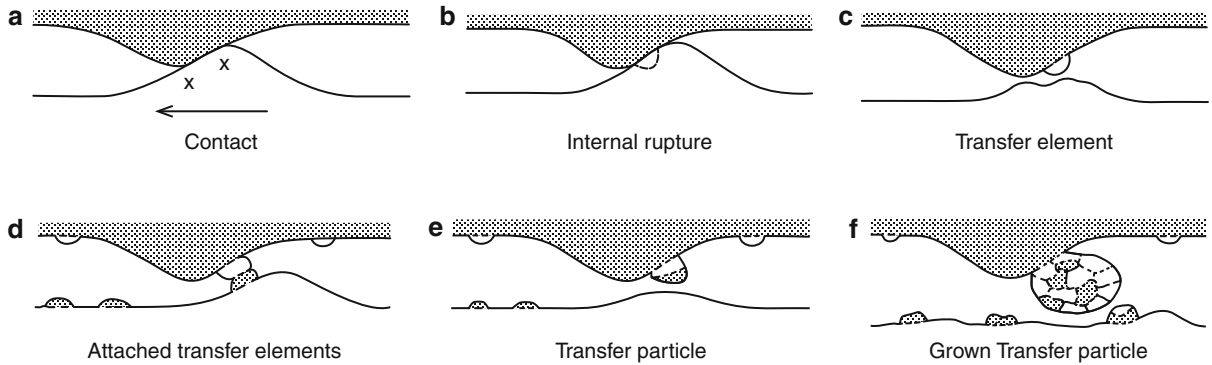
According to Clark et al.'s work, brush wear is affected by the stress of real contact area, and occurs like fatigue. Repeated stress causes large carbon particles to drop off from the brush surface.

Brush wear is often expressed by wear rate, that is, wear volume per unit sliding distance.

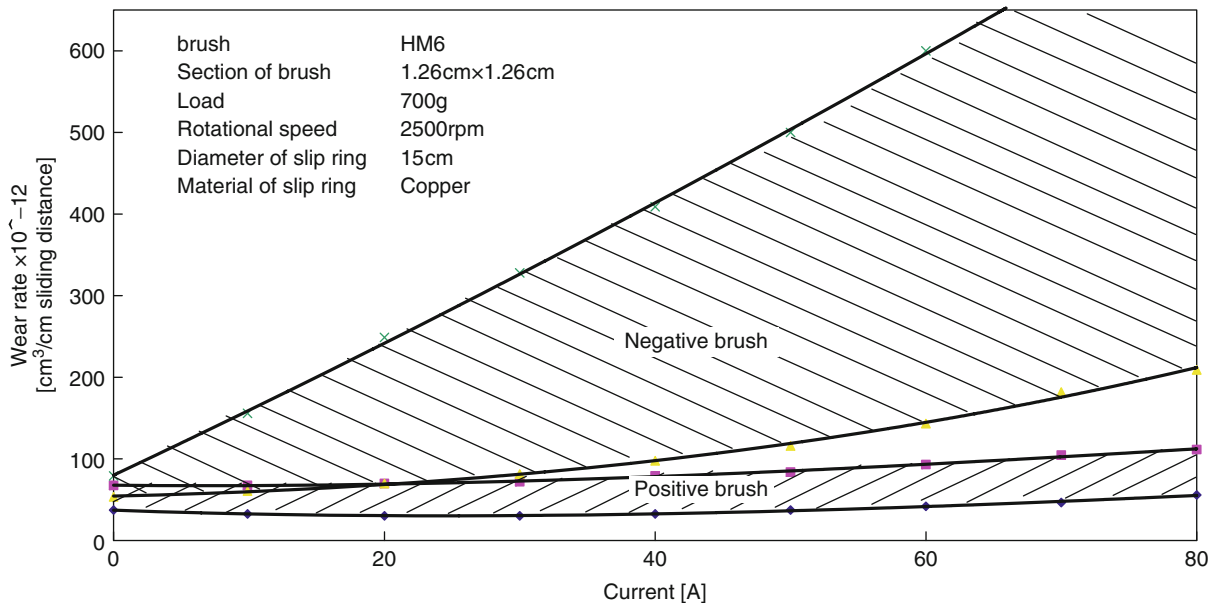
Holm proposed an equation of wear under the conditions of current and arc discharge:

$$W = P \left[W_0 + C_1 \{ \tau(2) + 2\tau(5) \} I + g \sqrt{Q} \right] + \omega Q \text{ cm}^3/\text{km}, \quad (8)$$

where P : mechanical load on a brush, I : current per brush, Q : electric charge transported by arcs during 1 km of sliding, ωQ : volume evaporated from the brush under



Sliding Electrical Contact Wear, Fig. 3 Transfer and growth wear model



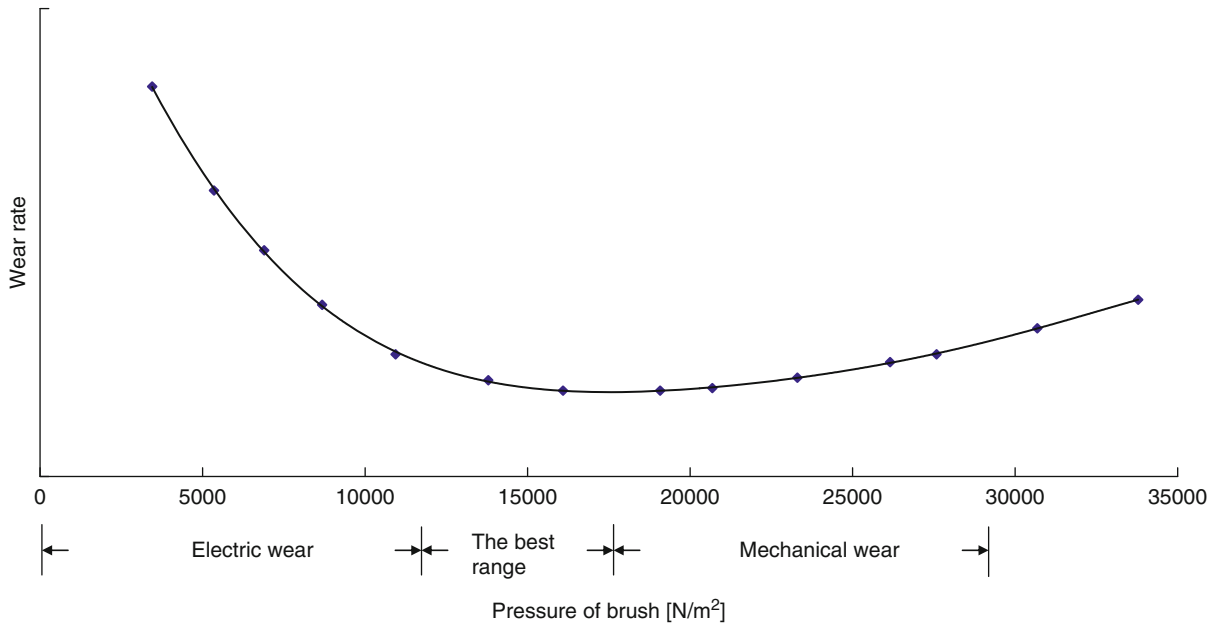
Sliding Electrical Contact Wear, Fig. 4 Wear rate vs. current

the influence of the arcs (Lancaster 1963; Lawson and Dow 1985), $\tau(2)$, $\tau(5)$: fractions of the test time during which the voltage is over 2 and over 5 V, respectively, the numeral 2 before $\tau(5)$: weight to the wear compared with $\tau(2)$.

Each term means that PW_0 : wear without current, $PC_1\{\tau(2) + 2\tau(5)\}I$: wear due to surface roughening by the flashes, $Pg\sqrt{Q}$: wear due to surface roughening by arcing where the square root is based on earlier measurements, and g is a coefficient that is determined so as to provide optimum agreement between formula and measurements.

Effect of Current on the Wear

Wear with current is generally larger than one without current. The current concentrates into a-spots at the contact surface between brush and commutator/slip ring (Bryant et al. 1995). Due to the current concentration, the temperature of a-spots is much higher than the average temperature of the brush. In the case of porous binder carbon, heated parts are oxidized and become fragile (Bryant 1991; Quinn and Winer 1985). Eventually, bonded carbon or graphite particles are broken away by sliding motion.



Sliding Electrical Contact Wear, Fig. 5 Wear rate vs. brush pressure

Figure 4 shows the relation between brush wear and current for a copper slip ring. In the case of a carbon brush, the wear of a negative brush is greater than that of a positive brush.

On the other hand, Thompson et al. reported that the wear of a positive brush is greater than that of a negative one for a steel slip ring.

In the case of a copper collector (commutator or slip ring), it is not clear why the wear of a negative brush is greater than that of a positive one. Hessler found that the surface under a positive brush is smoother than that under a negative one according to the optical observation of the ring surface, and noted that this is the reason (Hessler 1937). Another reason mentioned was that the temperature of a cathode is higher than that of an anode for the arc discharge.

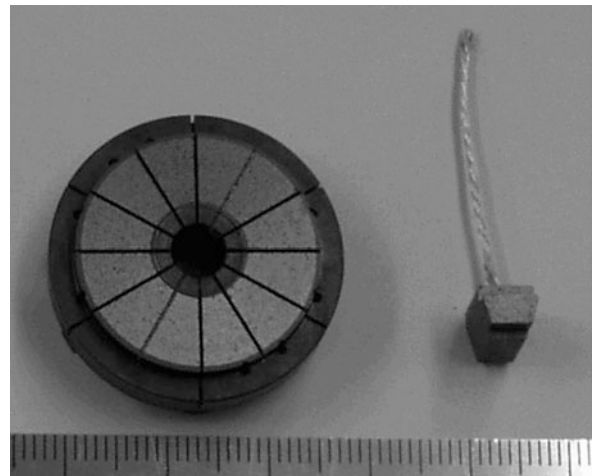
Effect of Brush Pressure on the Wear

Electric wear increases at too small a load, whereas mechanical wear increases at very large load. Figure 5 shows schematically the relation between wear rate and contact load (Lancaster 1962).

Key Applications

Automotive Applications

Figure 6 shows the commutator and brush of a DC motor driving an automotive fuel pump. Currently, most commutators of this type of motor are generally made of

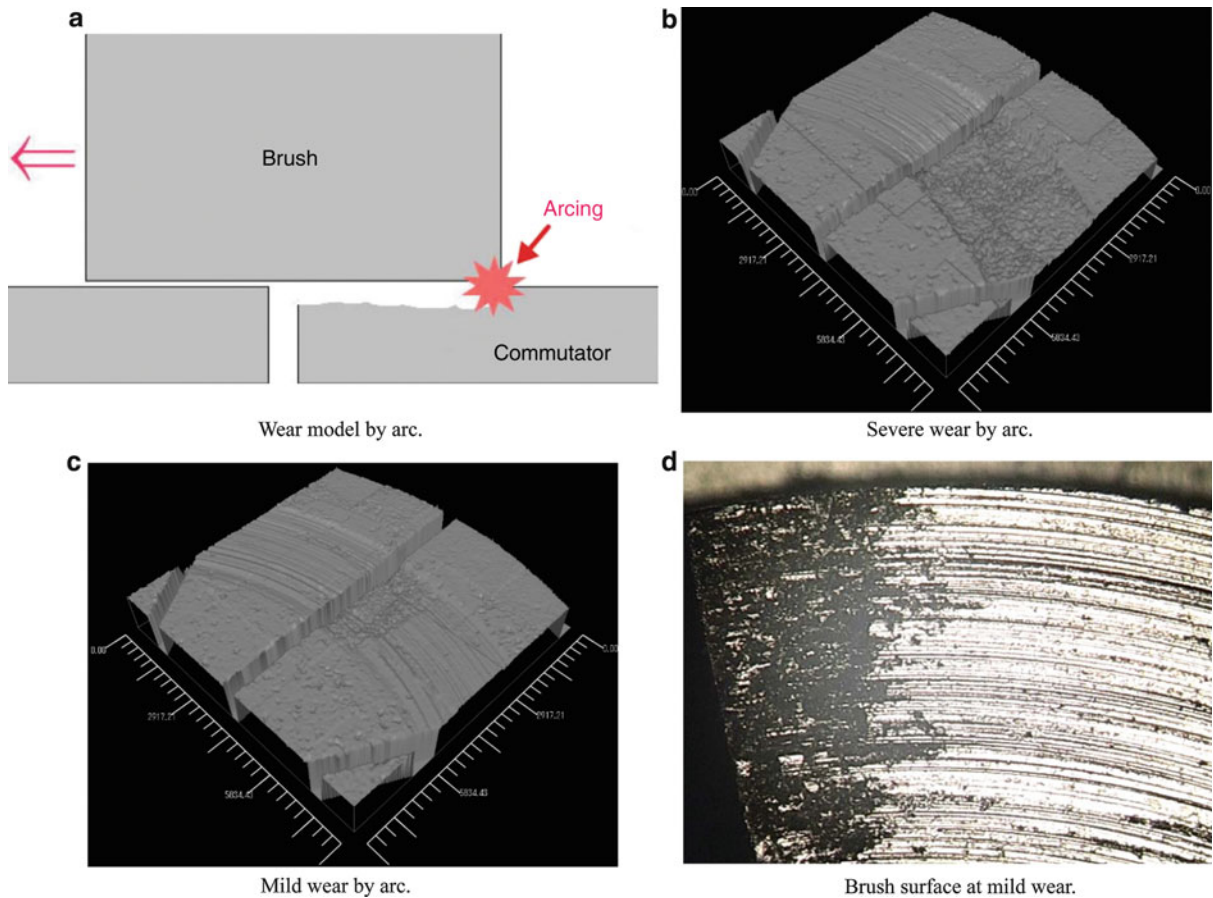


Sliding Electrical Contact Wear, Fig. 6 Wear rate vs. brush pressure

carbon and flat in their geometry. The commutator and brush are installed inside a gasoline tank and gasoline flows inside the motor.

Here, commutation is carried out in gasoline. In case of arcing, direct wear due to the arc is shown as Fig. 7a.

Commutator and brush are worn by the arc generated at the trailing edge of the brush. Simultaneously, mechanical wear is proceeding at the sliding surface due to sliding motion (Sawa and Shimoda 1992; Yamamoto et al. 1995).



Sliding Electrical Contact Wear, Fig. 7 Wear behaviors under arcing

If the wear by arc is severe, the eroded part by arc gradually extends. Finally, the eroded part reaches the thickness of a brush, as shown in Fig. 7b of a laser microscopic figure, and sliding condition becomes extremely bad, with the brush falling into the eroded hollow. This signifies the end of the part's lifetime.

On the other hand, when arc energy is small and the wear by arc is mild, the wear by arc and by sliding motion are balanced and the part eroded by arc does not extend over a certain amount, as shown in Fig. 7c. Figure 7d shows the brush surface in this case. The eroded part on the brush does not extend, as does the commutator surface.

Cross-References

- [Contacts Considering Adhesion](#)
- [Fatigue](#)
- [Surface Free Energy](#)
- [Temperature Effect on Fatigue](#)

References

- F.P. Bowden, D.T. Tabor, *The Friction and Lubrication of Solids* (Clarendon, Oxford, 1954)
- M.D. Bryant, A particle ejection mechanism for brush wear. *IEEE Trans. Compon. Hybrids Manuf. Technol.* **14**, 71–77 (1991)
- M.D. Bryant, A. Tewari, J.W. Lin, Wear rate reduction in carbon brushes, conducting current and sliding against wavy copper surfaces. *IEEE Trans. Compon. Hybrids Manuf. Technol.* **18**, 375–381 (1995)
- V.P. Hessler, Abrasion- A factor in electrical brush wear. *Electr. Eng. AIEE Trans.* **56**(8), 130 (1937)
- R. Holm, *Electric Contacts* (Springer, New York, 1967)
- J.K. Lancaster, The influence of the conditions of sliding on the wear of electrographitic brushes. *Br. J. Appl. Phys.* **13**, 468–477 (1962)
- J.K. Lancaster, The influence of arcing on the wear of carbon brushes on copper. *Wear* **6**, 341–352 (1963)
- D.K. Lawson, T.A. Dow, The sparking and wear of high current density electrical brushes. *Wear* **102**, 105–125 (1985)
- T.F.J. Quinn, W.O. Winer, The thermal aspects of oxidative wear. *Wear* **102**, 67–80 (1985)
- E. Rabinowicz, *Friction and Wear of Materials* (Wiley, New York, 1965)

- T. Sasada, S. Norose, H. Mishina, The behavior of adhered fragments interposed between sliding surface and the formation process of wear particles. *Trans. ASME J. Lubr. Technol.* **103** (1981)
- K. Sawa, N. Shimoda, A study of commutation arcs of DC motors for automotive fuel pumps. *IEEE Trans. Compon. Hybrids Manuf. Technol.* **15**, 193–197 (1992)
- P.G. Slade, *Electrical Contacts* (Marcel Dekker, New York, 1999)
- T. Yamamoto, K. Bekki, K. Sawa, A study on brush wear under commutation arc in gasoline, in *Proceedings of the 41th IEEE Holm Conference on Electrical Contacts*, 1995, pp. 323–329
- H. Zaidi, D. Paulmier, J. Lepage, The influence of the environment on the friction and wear of graphite carbons. *Appl. Surf. Sci.* **44**, 221–233 (1990)

Sliding Electrical Contacts and Materials

KOICHIRO SAWA

Department of System Design Engineering, Keio University, Kohoku-ku, Yokohama, Japan

Definition

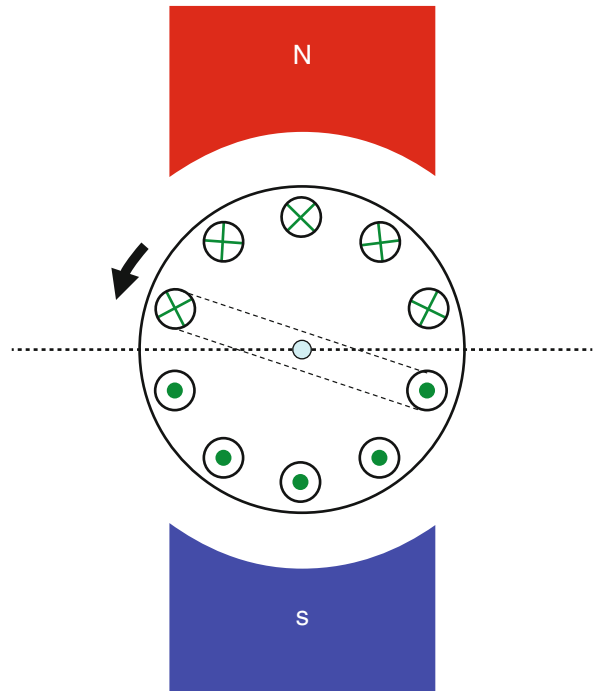
A sliding contact is an electrical contact where current or signal flows through the contact. The contacts consist of a brush (usually stationary part) and a slip-ring/commutator (rotating part). Low, stable contact voltage and low wear rate are required for materials used in sliding electrical contacts. Copper and copper alloy are mainly used as slip-ring or commutator for comparatively large current. For small current applications, precious metals also are used.

Scientific Fundamentals

Sliding electrical contacts usually consist of brush and slip ring/commutator (for more on brushes, see ► [Electrical Brushes](#)). This essay mainly describes commutators that carry out current commutation with brushes in DC motors, and slip rings that transfer current or signal between stationary part and moving part (Wilsdorf 1991; Liu 2001).

Commutators

In DC motors, the direction of current flow has to be controlled to realize a continuous rotation. Namely, when a unit coil passes through a magnetic neutral zone in [Fig. 1](#), the current direction of the coil has to be reversed to make the generated torque in the same direction. The reversal of current direction is called “commutation” in



Sliding Electrical Contacts and Materials, Fig. 1 Coil current distribution of dc motor

DC motors. The commutation is usually carried out with brush and commutator (Holm 1967; Shobert 1993).

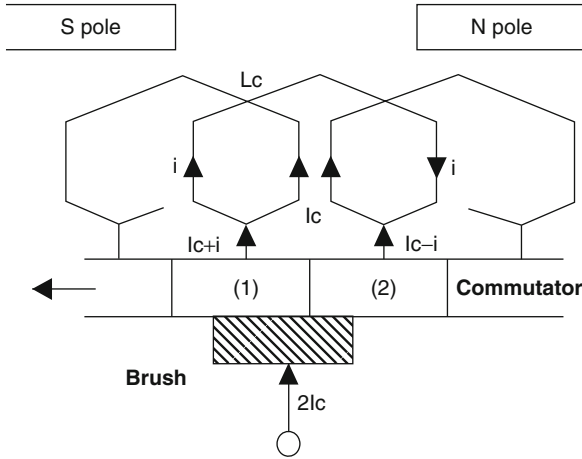
Usually, a commutator is cylindrical in shape, and coils are cylindrically placed. In [Fig. 2](#) they are horizontally spread. The commutator consists of commutator segments insulated from each other. Both terminals of each coil are connected with adjacent segments.

A coil just passing through the neutral zone is focused and called a “commutation coil.” Both terminals are connected with segment 1 and 2. When the brush is on segment 1, the current of the commutation coil flows clockwise. If coils and commutator move in the left direction, the brush will touch segment 2, and the direction of the coil current will change to counterclockwise. The commutation is carried out.

The current changes with time during the commutation, based on the following equation (Holm 1958):

$$L_c \frac{di}{dt} + R_1(I_c + i) - R_2(I_c - i) = 0, \quad (1)$$

where R_0 is defined as the contact resistance when the brush contacts the segment entirely (Shobert 1954) and then the resistance R_1 between the brush and the segment 1 and R_2 between the brush and segment 2 are expressed by (2).



Sliding Electrical Contacts and Materials, Fig. 2
Commutation circuit of dc motor

$$\begin{cases} R_1 = \frac{R_0}{1-t/T_c} \\ R_2 = \frac{R_0}{t/T_c} \end{cases}, \quad (2)$$

where T_c : commutation period and t : time.

The voltage between brush and segment 1 v_t is

$$v_t = (I_c + i)R_1 \quad (3)$$

Equations (1) and (3) are normalized using I_c to the current and T_c to the time, and Equations (5) and (6) can be obtained.

$$\frac{d(i/I_c)}{d(t/T_c)} + \left(\frac{R_0 T_c}{L_c} \right) \left(\frac{1+i/I_c}{1-t/T_c} + \frac{i/I_c-1}{t/T_c} \right) = 0 \quad (4)$$

$$\frac{d\theta}{d\tau} + \rho \left(\frac{1+\theta}{1-\tau} + \frac{\theta-1}{\tau} \right) = 0 \quad (5)$$

$$v_t = \frac{v_s}{2I_c R_0} = \frac{1+\theta}{2(1-\tau)} \quad (6)$$

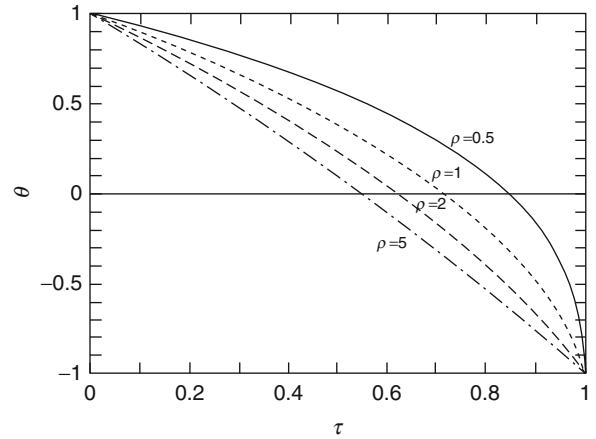
Initial condition: $\tau = 0$, $\theta = 1$,

where $\theta = i/I_c$, $\tau = t/T_c$, $\rho = \frac{R_0 T_c}{L_c}$.

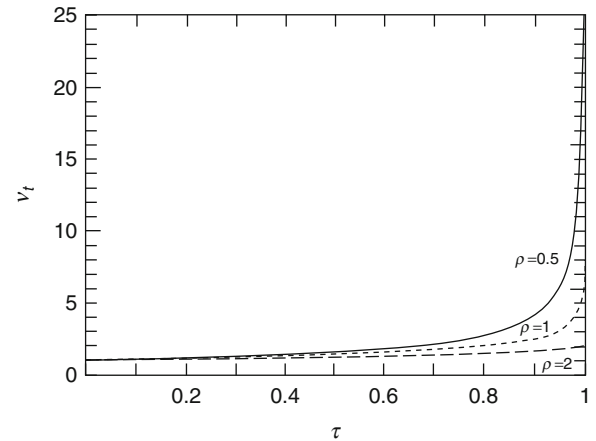
Figure 3 shows the solution of (5) with parameters of ρ . In the case that ρ is small, the commutation current is delayed in change and it makes difficult commutation. Figure 4 shows the voltage of brush trailing edge v_t and in case of $\rho \ll 1$ the voltage v_t becomes theoretically very high at the end of the commutation.

Commutation Arc

As mentioned above, the voltage of brush trailing edge v_t becomes very high. However, it is well known that there is



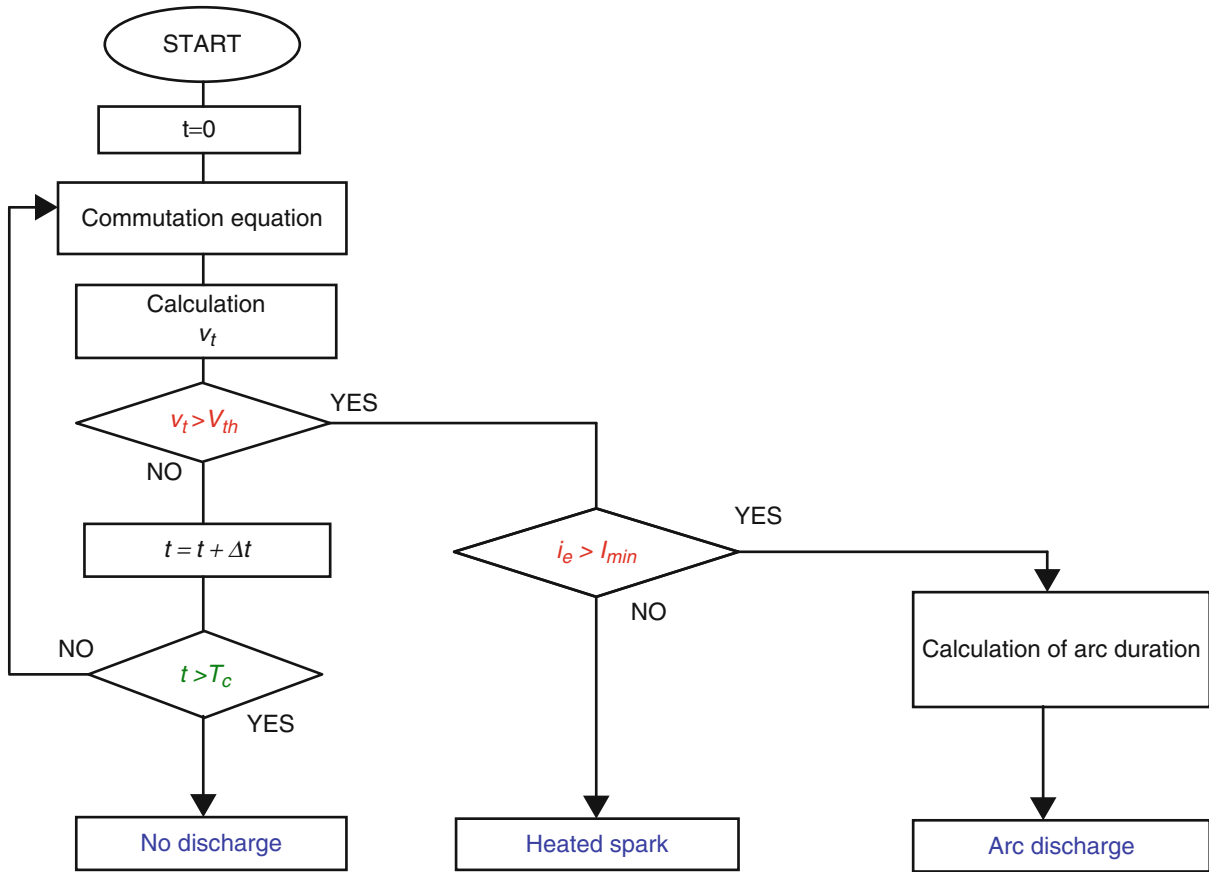
Sliding Electrical Contacts and Materials, Fig. 3 Change of
commutation current



Sliding Electrical Contacts and Materials, Fig. 4 Change of
trailing voltage drop

a relation between contact voltage and highest temperature of the contact ($\Phi - \theta$ relation) so that the highest temperature may exceed a threshold temperature of contact materials (boiling temperature for metals and sublimation temperature for carbon) (Holm 1967; Slade 1999). Consequently, metallic contact will rupture around the threshold temperature. The voltage corresponding to the threshold temperature is expressed as V_{th} .

Therefore, when the brush contact voltage reaches V_{th} , actual metallic contact is broken down and the commutation is over before a geometrical commutation period T_c . In some cases, the current between brush and segment 1 ($I_c + i$) is not zero at actual end of the commutation and



Sliding Electrical Contacts and Materials, Fig. 5 Calculation model of commutation arc

that value of $(I_c + i)$ is called “residual current I_e .” If the residual current is larger than the minimum arc current of a contact material, arc discharge will occur.

Figure 5 shows a diagram of the arc occurrence model explained above. The arc duration τ_a is approximately expressed by the following equation (Takaoka et al. 2001):

$$\tau_a = \frac{L_c}{v_a} (I_e - I_{\min}) \quad (7)$$

where L_c is inductance of the commutation coil and v_a is arc voltage.

Commutator Materials

In small and large DC motors, silver bearing copper is usually used as a main material of the commutator. Silver content is generally about 1%.

For micro DC motors, precious metal alloys, gold alloy, and silver alloy are used (see Table 1).

In addition, a carbon commutator is used for a special application, a DC motor driving an automotive fuel

pump, because a stable operation and long life can be expected in various kinds of gasoline.

Slip Ring Materials

Copper alloys are used for large current slip rings because of minimum heating.

Low contact resistance and resistance variation are required for various slip-ring applications (Glossbrenner and Sun 1963). Silver and its alloys are effective in cost comparing to gold or platinum, because silver is resistive to form hard oxide and the alloys are resistive to sulfation (Smith 1993). Further, silver plating and clad are more effective in cost, but their thickness is important to wear life.

For very low voltage and current, for example, much less than 1 V or 1 A, gold or gold alloy is used as a slip ring material. As it is well known, gold is very inert to reaction with atmospheric gases, not from oxide and sulfide. So, it is also used to protect contact surface in case of long-term idling or very few operations. In addition, gold has low

Sliding Electrical Contacts and Materials, Table 1 Precious metals for slip-ring and commutator

	Composition (%)	Melting point (°C)	Hardness (HV)	Electric conductivity LACS (%)	Density (g/cm ³)	Applications
Au alloy	Au-Ag8	1,058	30	28.7	18.0	Commutator
	Au-Ag10	1,055	30	25.4	17.9	
	Au-Ag20	1,045	33	18.1	16.6	
	Au-Ag25	1,040	35	16.6	16.0	
	Au-Ag40	1,005	40	15.6	14.5	
	Au-Ag90	970	29	48	11.0	
Pt alloy	Pt-Pd10	1,550	90	6.2	19.9	Slip ring
	Pt-Pd20	1,560	110	5.7	18.6	
Pd alloy	Pd-Cu15	1,380	100	4.6	11.2	Potentiometer
Ag alloy	Ag-Cu7.5	799	56	90	10.4	Commutator for micro motor
	Ag-Cu10	778	62	86	10.3	
	Ag-Cu90	778	60	80	9.1	
	Ag-Cu6-Cd2	880	65	43	10.4	Commutator for micro motor
	Ag-Cu24.5-Ni0.	810	135	68		
	Ag-In18	746	50			Commutator
	Ag-Cdl	959	35	92	10.5	

catalytic activities to organic gases, while platinum and palladium group form much friction polymer in organic gases during operations. Generally, gold is used as a cladding or electroplate on a bulk metal to reduce the cost, but its thickness should be properly selected from wear life.

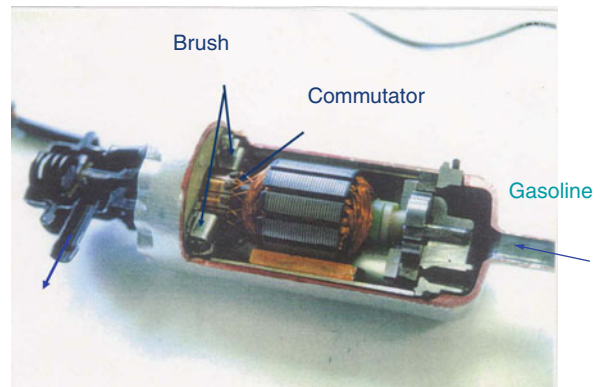
Key Applications

DC Motor Applications

DC motors from micro size to large size are widely used in various applications (Shobert and Diehl 1955). A motor used for an automotive application is presented to introduce fundamental phenomena of the DC motor. Figure 6 shows an inside of a small DC motor driving an automotive fuel pump. In this motor, gasoline or fuel flows inside the motor and the commutation is carried out in fuel (Takaoka et al. 2001).

Figure 7 shows an equivalent commutation circuit. The circuit was originally proposed by Olney, where a residual current, arc voltage, and arc duration are easily measured (Olney 1950).

Figure 8 shows typical voltage and current waveform of the commutation current. The final stage of the commutation is enlarged, and the voltage between brush and commutator segment abruptly increases and an arc



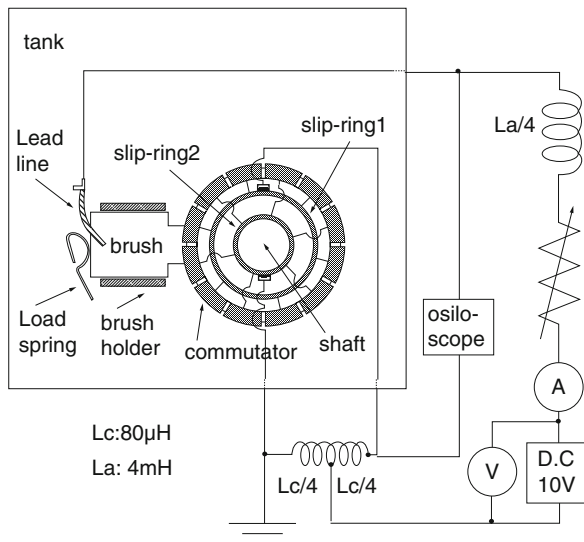
Sliding Electrical Contacts and Materials, Fig. 6 Photo of dc motor driving a fuel pump

discharge takes place when metallic contact between brush and commutator segment ruptures. The commutation current that remains at the breakdown of the metallic contact is called a residual current, as mentioned before. If the residual current is larger than the minimum arcing current, arc discharge occurs. In the experiment the residual current and arc duration are measured.

Experimental results between arc duration and residual current are shown in Fig. 9. There are comparatively

large fluctuations, but a linear relation between arc duration and residual current is confirmed, as expected by (7).

Figure 10 shows photos of commutator surfaces after 50-h wear tests in gasoline for carbon brush and carbon fiber brush. The material formed on the commutator surface in gasoline is not oxide film, but carbon film transferred from the brush material. The carbon film gets thick with the current. Particularly thick



Sliding Electrical Contacts and Materials, Fig. 7 Equivalent commutation circuit

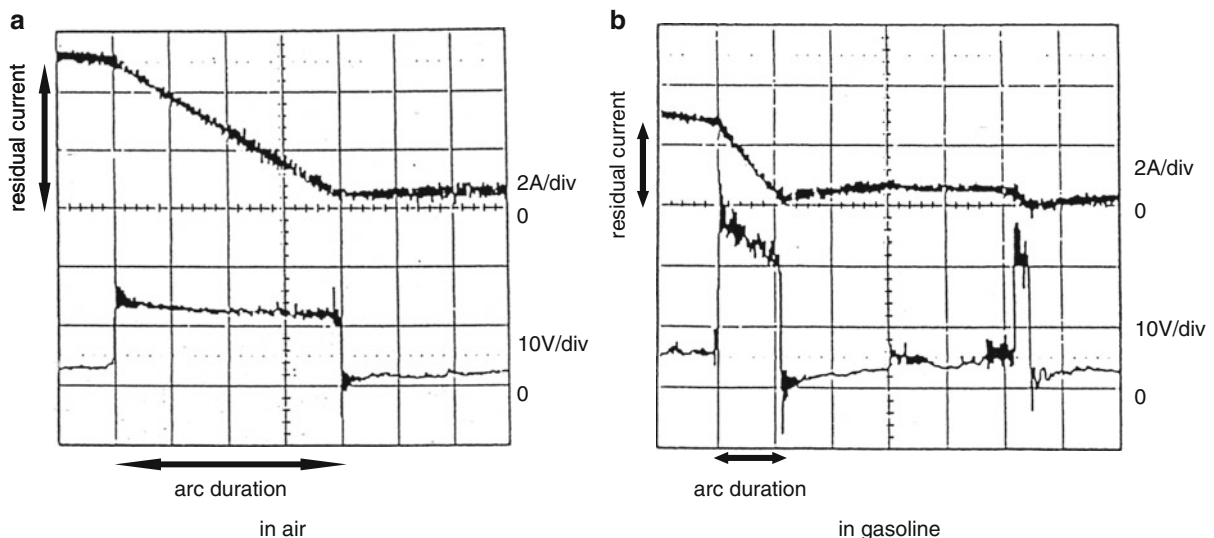
carbon film formed on the surface causes severe wear. The carbon fiber brush makes thinner carbon film on the commutator than the carbon brush.

Figure 11 shows photos of the brush surface after a 50-h test in gasoline. For a carbon brush, the sliding surface is completely different from that in air. The surface in air is comparatively smooth, but in gasoline arc traces are scattered on almost the entire surface. In contrast to carbon brush, arc traces are observed only on the trailing part of carbon fiber brush except for the case of 10 A. At the current level of 10 A, even a carbon fiber brush surface is covered with arc traces. The distribution of arc traces seems to affect the sliding condition and brush wear.

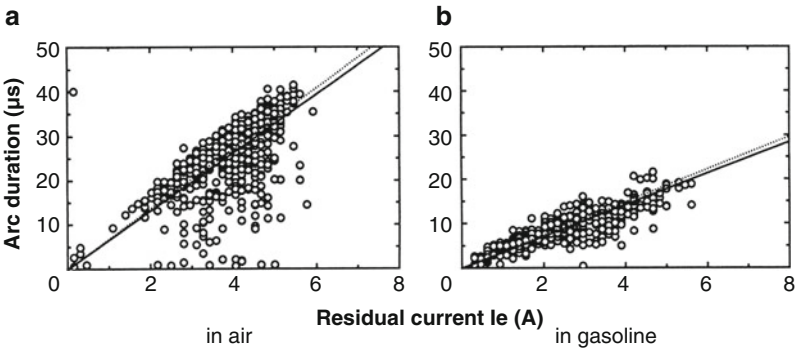
Slip Ring Applications

There are many applications of slip rings (Dow 1982; Glossbrenner and Sun 1963; Pentlicki and Glossbrenner 1971). As an example, the sliding quality and the deterioration process of Au plated slip-ring and Ag-Pd brush system is presented, which is used mainly in the chip-mounter and others (Kobayashi et al. 2007). A slip ring is hard-gold plated on Ni base plating on a brass ring as shown in Fig. 12, while a brush consists of ten Ag-Pd alloy wires and two brushes are mated with one ring.

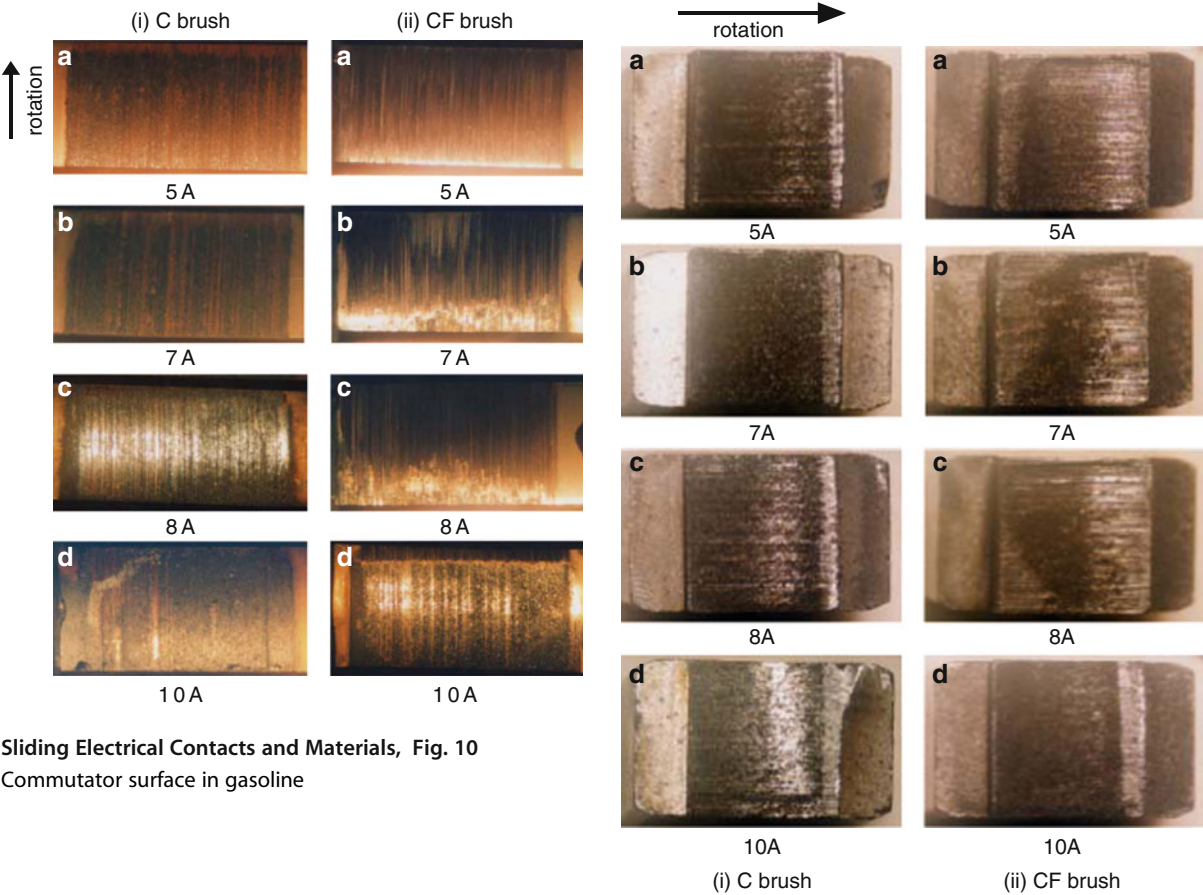
This slip ring system consists of eight rings, and two rings are used in one circuit, as shown in Fig. 13. So, four pairs of rings are subjected to the test; two pairs with standard amount of lubricant and two with a double amount. Typical contact voltage change with time is



Sliding Electrical Contacts and Materials, Fig. 8 Voltage and current waveform at final stage of commutation



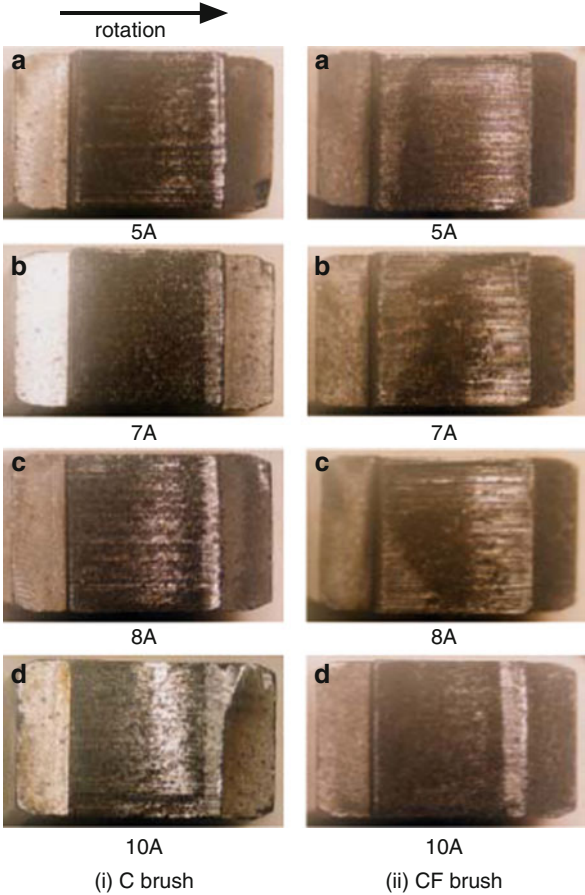
Sliding Electrical Contacts and Materials, Fig. 9 Arc duration τ_a vs. residual current I_e



Sliding Electrical Contacts and Materials, Fig. 10
Commutator surface in gasoline

shown in Fig. 14. Contact voltage drop rises gradually as time passes with small voltage fluctuations.

The pair of No.3–4, with more lubricant, reached the end of lifetime after 3,782 h. But other pairs achieved long life, independent of the amount of lubricant. This result shows that any amount of lubricant over a threshold value below which the lubricant is not effective may not affect the lifetime.



Sliding Electrical Contacts and Materials, Fig. 11
Brush surface in gasoline

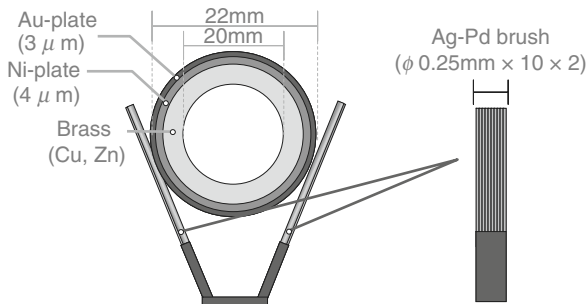
Figure 15 shows surface images of slip-ring by optical microscope (x75) at several operation times. It presents a typical deterioration process as follows:

- (a) Some worn tracks appeared.
- (b) Surface color changes partly into brown, which looks like a mixture of lubricant and wear particle.
- (c)–(d) Wear progressed slowly with time and tracks became clear.
- (e) Brown parts gradually disappeared and Ni plating layer or base metal (white tracks on the image) may be exposed.
- (f) Almost all tracks of base metal were oxidized (the end of lifetime).

In general, if one track is oxidized, the current moves to other tracks, and it will increase contact temperature. Lubricant is accelerated to volatilize and wear is also accelerated.

After the test, element distribution on the surface of a slip-ring was analyzed by EPMA.

Figure 16 shows an example of surface element analysis at the ring reaching lifetime. Au and Ni plating layers have worn out completely and underlying metal (brass including Zn) is exposed. In addition, a carbon element



Sliding Electrical Contacts and Materials, Fig. 12 Slip-ring and brush system

is not detected, which means that almost all lubricant disappears. The ring surface was oxidized completely.

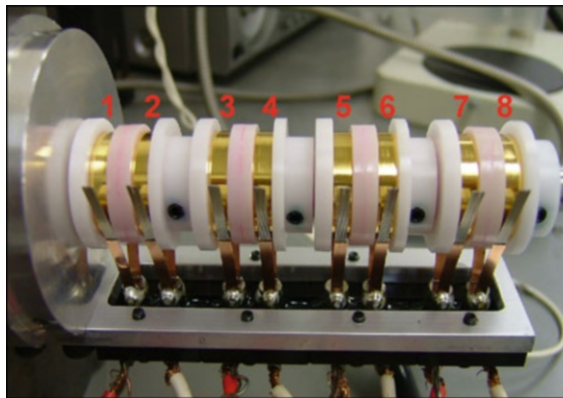
Lifetime is dependent on load current and other factors. However, a typical deterioration process can be described as follows; When protected layers like Au and Ni plate are worn out, a base material of the ring, for example, bronze is exposed and the surface starts to be oxidized. At this stage the contact resistance rapidly increases as mentioned above and then the lifetime is over. The lifetime is decided mainly by the surface wear rate. However, the problem is that the lifetime varies widely at each operation even with similar operation conditions.

Applications Accompanying with Arc Discharge

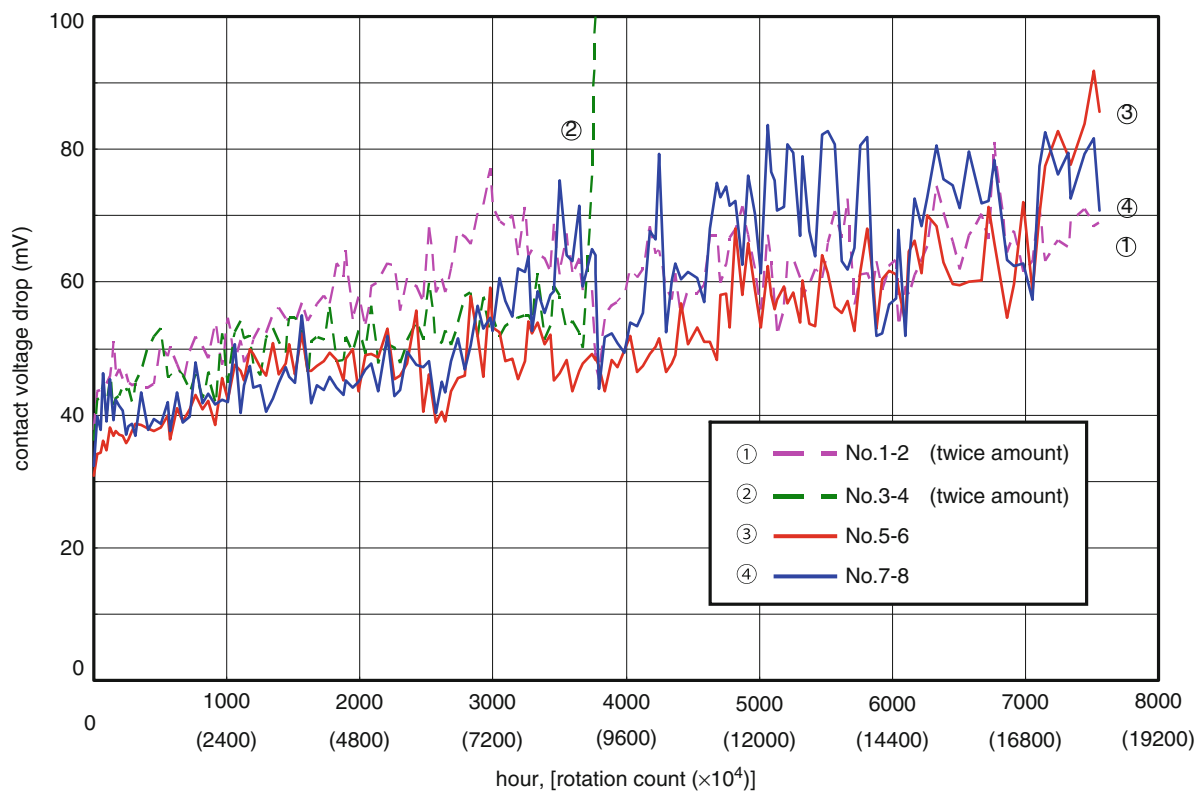
Metals and alloys have to be resistive to high temperature in such applications. However, tungsten or metal oxide additives are not used in sliding contacts, unlike relays, breakers, and similar switches. Such additives usually increase abrasiveness of contacts and then greatly decrease wear life of sliding contacts. Suppress of arc discharge in sliding contacts is often realized by mechanical and electrical design rather than material selection (Holm 1967; Slade 1999).

Potentiometer Application

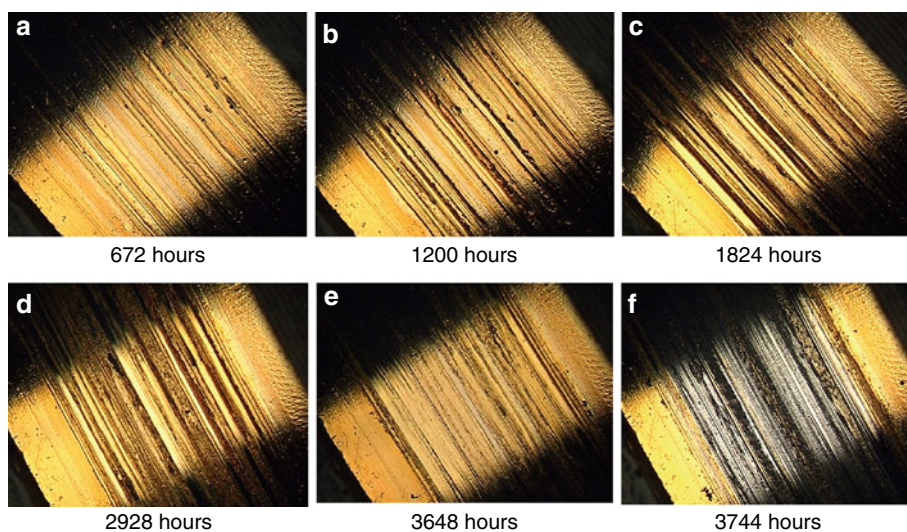
A resistor material is required to provide a desired resistance gradient as the brush passes across it (Slade 1999). Wire-wound potentiometers use fine precious metal wires with resistivity and diameter to obtain the correct resistance gradient. Similarly, film potentiometers use a metal or graphite film of proper resistivity, thickness, and width to obtain the desired resistance distribution. Generally, a resistor material is an alloy of platinum group with



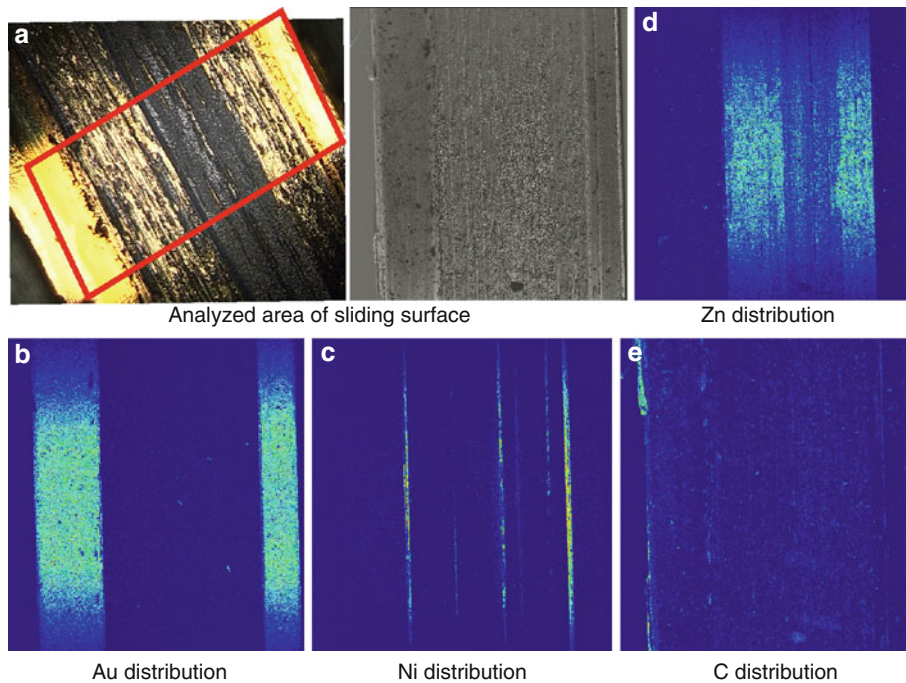
Sliding Electrical Contacts and Materials, Fig. 13 Overview of a tested slip-ring system



Sliding Electrical Contacts and Materials, Fig. 14 Contact voltage drop in sliding condition



Sliding Electrical Contacts and Materials, Fig. 15 Surface change of slip-ring of No. 4



Sliding Electrical Contacts and Materials, Fig. 16 Element distribution of ring at the end of lifetime

high resistivity. Graphite film is a mixture of graphite and other conductive material to provide a desired resistivity.

Cross-References

- [Brush Materials](#)
- [Contact Boiling Voltage](#)
- [Contact Melting Voltage](#)
- [Electrical Brushes](#)
- [Electroplating](#)
- [Sliding Electrical Contact Wear](#)
- [Surface Roughness](#)

References

- T.A. Dow, Thermomechanical effects in high current density electrical slip rings. *Wear* **79**, 93–105 (1982)
- D. Frank, Olney: “resistance commutation”. *AIEE Trans.* **69**, 1207–1218 (1950)
- E.W. Glossbrenner, J.K. Sun, Effects of parameters in miniature sliding contacts. Engineering Seminar on Electrical Contacts, University of Maine, 1963, Paper 5
- R. Holm, *Electric Contacts* (Springer, New York, 1967)
- R. Holm, Contribution to the theory of commutation on DC machines. *AIEE Trans. Power Appar. Syst.* **77**, 1124–1129 (1958)
- T. Kobayashi, K. Sawa, K. Endo, G. Ou, H. Hagino, A study of high reliability system of Au-plated slip-ring and Ag-Pd brush for power supply, in *Proceedings of the 53th IEEE Holm Conference on Electrical Contacts*, Pittsburgh, 2007, pp. 194–199
- H.P. Liu, R.W. Carnes Jr., J.H. Gully, Measurement and prediction of brush interface temperature at sliding electrical contacts, in *Proceedings of the 38th IEEE Holm Conference on Electrical Contacts*, Philadelphia, 1992, pp. 143–148
- C.J. Pentlicki, E.W. Glossbrenner, The testing of Slip rings and brushes for space applications, in *Electric Contacts 1971, Proceedings of the Holm Conference on Electrical Contacts* (Illinois Institute of Technology, Chicago, 1971), pp. 157–172
- E.I. Shobert II, Electrical resistance of carbon brushes on copper rings. *IEEE Trans. Power Appar. Syst.* **13**, 788–797 (1954)
- E.I. Shobert II, Sliding electrical contacts, in *Proceedings of the 39th IEEE Holm Conference on Electrical Contacts*, Pittsburgh, 1993, pp.123–134
- E.I. Shobert II, J.E. Diehl, A new method of investigating commutation as applied to automotive generators. *Power Appar. Syst.* **16**, 1592–1603 (1955)
- P.G. Slade, *Electrical Contacts* (Marcel Dekker, New York, 1999)
- E.F. Smith et al., Screening of contact materials for low speed slip ring assemblies, in *Electrical Contacts 1993, Proceedings of the IEEE-Holm Conference* (Institute of Electrical and Electronics Engineers, Pittsburgh, 1993), pp. 157–170
- M. Takaoka, T. Aso, K. Sawa, A commutation performance and wear of carbon-fiber brush in gasoline, in *Proceedings of the 47th IEEE Holm Conference on Electrical Contacts*, Montreal, 2001, pp. 44–49
- D.K. Wilsdorf, Uses of theory in the design of sliding electrical contacts, in *Proceedings of the 37th IEEE Holm Conference on Electrical Contacts*, Chicago, 1991, pp. 1–24

Sliding Resistance

- [Friction \(Concepts\)](#)

Sliding Velocity

► Gear Sliding

Sliding Wear

► Wear in Gears

Sliding Wear in Mixed EHL

ASHLIE MARTINI

School of Engineering, University of California Merced,
Merced, CA, USA

Synonyms

Sliding wear in mixed elastohydrodynamic lubrication; Sliding wear in mixed lubrication; Sliding wear in partial EHL; Sliding wear in partial elastohydrodynamic lubrication

Definition

Wear may occur in mixed elastohydrodynamic lubrication (EHL) interfaces at the areas of direct solid contact. This is an extremely complicated process because wear is dependent on local interface conditions, which themselves can be significantly affected by wear.

Scientific Fundamentals

Introduction

Understanding wear in mixed EHL contacts is challenging for two primary reasons. First, the amount of wear is directly related to local lubrication effectiveness and contact severity, which are difficult to quantify in mixed EHL. The direct result is that effective models for mixed EHL wear need to incorporate the effects of hydrodynamic flow, elastic deformation of the contacting bodies, surface roughness and topography, and changes of viscosity and density with pressure, as well as the wear itself. Simple models with one or two stochastic parameters may not be sufficient to describe these integrated and complicated phenomena. The second difficulty is the mutual dependency of wear and mixed lubrication characteristics. Wear changes the surface topography in real time. The surface

topography may, in turn, significantly affect mixed lubrication characteristics and contact severity. Contact severity may then greatly influence the wear process. This interdependency is difficult to describe with simplified time-independent models, and a satisfactory solution may require a transient deterministic simulation. Wear predictions from one example of such a simulation is shown in Fig. 1.

Wear Models

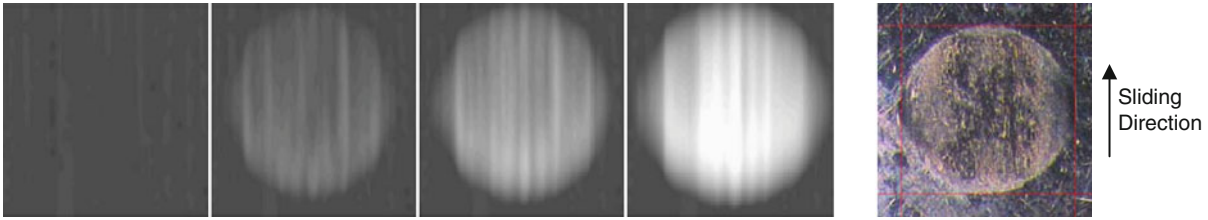
Wear takes place at locations in the interface where surfaces are in direct contact. The rate of material removal is directly correlated to “contact severity,” which is a complicated concept that may involve contact pressure, sliding speed, asperity deformation, interfacial temperature, friction, subsurface stresses, and some other parameters. In general, however, it is agreed that wear is a function of material properties, operating conditions, and local characteristics of an interface. Many different theoretical and empirical models have been developed to define this relationship for different applications. In fact, in 1995 Meng and Ludema (1995) reviewed 35 years of the journal *Wear* and 14 years of proceedings from the “Wear of Materials” conferences and identified 182 distinct wear equations. However, the form of many of the most common form of these expressions is

$$\frac{dW}{dt} = k \frac{p^\alpha u^\beta}{H^\gamma} \quad (1)$$

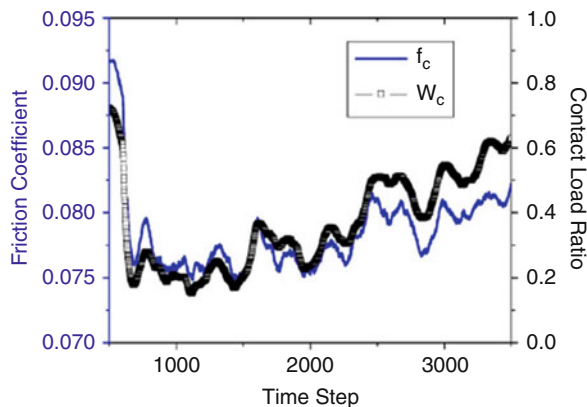
where dW/dt is wear rate, p is contact pressure, u is relative sliding velocity, and H is material hardness. In (1), k is the wear coefficient and the three exponential constants, α , β , and γ , differ based on which wear law is being used. One of the most well-known laws of this form (likely because of its simplicity) is Archard’s wear law, where $\alpha = \beta = \gamma = 1$. However, many more models of similar form are available, some of which are summarized in Goryachev (1998).

Phases of Wear

A typical system may be subject to wear in three phases: running in, steady-state, and accelerated or catastrophic wear. The running-in phase occurs at the onset of sliding and is associated with a high wear rate. During this phase, tall asperity peaks are quickly removed and the surfaces become smoother. After the tallest peaks are removed, the wear rate decreases and gradually stabilizes to the steady state. In the steady-state phase, the wear rate is relatively stable, and it is in this phase that it is desirable for machine components to operate for as long as possible. Lastly, in some cases, steady state gives way to catastrophic wear in which the wear rate once again increases. This is a general



Sliding Wear in Mixed EHL, Fig. 1 Evolution of a wear scar on the ball during a ball-on-disk experiment operating in mixed EHL (Zhu et al. 2007). *Left* figures from a full numerical mixed EHL simulation, *right* figure from experiment



Sliding Wear in Mixed EHL, Fig. 2 Simulation-predicted friction coefficient (solid line) and contact load ratio (squares) as functions of time for the ball-on-disk case illustrating the relationship between friction and contact area during the wear process (Zhu et al. 2007)

description of the wear process; actual observations may vary. For example, the running-in may be too short to observe, or a system may not undergo catastrophic wear even after a very long time.

Relationship Between Friction and Wear

The phases of wear that a system goes through can, in some cases, be correlated to the change in the friction coefficient over time. It has been observed via mixed EHL simulation (Zhu et al. 2007) that before significant wear occurs, the friction coefficient is quite high due to severe asperity peak contacts. Then there is a sharp decrease of the friction coefficient. This corresponds to the running-in phase in which tall asperity peaks are worn away quickly and the surface undergoes mechanical polishing. After this initial drop-off, the friction gradually increases. These simulation predictions are consistent with experimentally observed trends in both dry and lubricated contacts.

Physically, friction and wear in mixed EHL are related through the real area of contact, or more specifically the contact load ratio (see Fig. 2). The wear process may increase or decrease the percentage of load supported by direct solid contact. Since contact friction is typically larger than hydrodynamic friction, this will affect the overall mixed EHL friction and in turn the wear behavior.

Key Applications

Mixed EHL is an important branch of lubrication theory that describes lubrication mechanisms in non-conformal contacts, which can be widely found in many mechanical components such as various gears, rolling bearings, cams and followers, hydraulic vane pumps, ball screws, traction drives and continuous variable transmissions, and metal rolling tools. Even components that operate in a full-fluid EHL regime during standard operation may experience periods of mixed EHL during start-up and slow-down. The ability to predict wear that may occur during mixed EHL is important for all such applications since wear will directly affect interface performance and long-term component durability.

References

- I.G. Goryachev, *Contact Mechanics in Tribology* (Kluwer, Dordrecht, 1998), pp. 191–201
- H.C. Meng, K.C. Ludema, Wear models and predictive equations: Their form and content. *Wear* **181–183**, 443–457 (1995)
- D. Zhu, A. Martini, W. Wang, Y.H. Hu, B. Lisowski, Q. Wang, Simulation of sliding wear in mixed lubrication. *ASME J. Tribol.* **129**, 544–552 (2007)

Sliding Wear in Mixed EHL Contacts

► [Wear in Gears](#)

Sliding Wear in Mixed Elastohydrodynamic Lubrication

- ▶ [Sliding Wear in Mixed EHL](#)

Sliding Wear in Mixed Lubrication

- ▶ [Sliding Wear in Mixed EHL](#)

Sliding Wear in Partial EHL

- ▶ [Sliding Wear in Mixed EHL](#)

Sliding Wear in Partial Elastohydrodynamic Lubrication

- ▶ [Sliding Wear in Mixed EHL](#)

Slip

- ▶ [Shear Localization, the Limiting Stress, and Other Forms of Liquid Failure](#)

Slip-Jaw Clutch

- ▶ [Overrunning Clutch](#)

Slips and Trips

- ▶ [Tribology in Daily Life: Footwear-Surface Interactions in Pedestrian Slips](#)

Slip-Wave Destabilization

- ▶ [Contacts Involving Wave Propagation](#)

SMAT – Surface Mechanical Attrition Treatment

- ▶ [Surface Nanocrystallization and Hardening \(SNH\)](#)

Smooth Spherical Inverse Filter Method (SSIF)

- ▶ [Elasticity Theory for Spherical Bearings](#)

S-N Curve

- ▶ [Stress-Life Theories](#)

S-N Diagram

- ▶ [Stress-Life Theories](#)

Soak Control

- ▶ [Polymers in Biotribology](#)

Solar Planetary

- ▶ [Epicyclic Gear Trains](#)

Solid Film Lubricants

- ▶ [Solid Lubricants](#)

Solid Lubricant Films Deposited by Burnishing

LEV RAPOPORT

Department of Sciences, Holon Institute of Technology,
Holon, Israel

Synonyms

Reduction of friction and wear by rubbing of solid lubricant particles into ground or textured surfaces

Definition

Burnishing is the process of creating solid lubricant films by rubbing of the solid lubricant powders into the surfaces.

Scientific Fundamentals

Metal dichalcogenites, MoS_2 and WS_2 , are well known for their solid lubricant properties (Bhushan and Gupta 1991). Solid lubricant films are widely used to decrease the friction coefficient in definite contact conditions. Low friction of contact pairs rubbed with platelet solid lubricant is attributed to shearing of the weak inter-layer, typically van der Waals (vdW) bonds in graphite, MoS_2 , or WS_2 films. In the past few years, inorganic fullerene-like (IF) supramolecules of metal dichalcogenide MX_2 ($\text{M} = \text{Mo}, \text{W}, \text{etc.}; \text{X} = \text{S}, \text{Se}$), materials with structures closely related to (nested) carbon fullerenes and nanotubes, have been synthesized (Tenne et al. 1992; Margulis et al. 1993). Experiments have shown that IF- WS_2 provides low friction and wear loss under different contact conditions with oil and grease, and impregnated into porous matrix (Rapoport et al. 1997, 1999, 2001; Cizaire et al. 2002; Shahar et al. 2010). Excellent tribological properties of IF nanoparticles under high contact pressure (>1 GPa) have been attributed to transfer of exfoliated thin sheets to the contact surfaces. The continuous supply of solid lubricant nanoparticles to the sliding contact is of great importance here. It has been concluded that formation of “pockets” on the rubbed surface provides a steady flow of lubricants and easy shearing of the metal–metal interface, even under extreme loads.

Advanced ceramics such as alumina, zirconia, silicon carbide, and nitride have growing potential for wear protection. When ceramic materials rub against another surface, scratching and damage to the mating surface lead to its rapid deterioration. Solid lubricant films on the surface of ceramics prevent direct contact between rubbed surfaces and thus decrease friction and wear (Fusaro 1982).

Surface texturing as a means of enhancing tribological properties of mechanical components has been well known for many years. Fundamental research work on various forms and shapes of surface texturing for tribological applications has been carried out by several research groups worldwide, and various texturing techniques are employed in these studies, including machining, ion beam texturing, etching, and laser texturing. Almost all of these are experimental in nature and most are based on the idea that surface texturing provides micro-reservoirs to enhance lubricant retention or micro-traps to capture wear debris. Laser surface texturing (LST) seems to be the most advanced of the methods of surface texturing for tribological applications (Etsion 2009). LST is starting to gain attention in the tribology community, as is evident from the growing literature on this subject. Indeed, LST provides substantial improvement of tribological performance under friction with fluid lubricant. The geometrical parameters of LST have been optimized for fluid lubrication of flat surfaces under different contact conditions.

In recent years, laser texturing has been combined with incorporation of solid lubricant into micro-reservoirs. Solid lubricant that is stored in the dimples can be released to the interface, thus increasing the longevity of rubbed surfaces. However, friction and wear of solid lubricant films on LST steel surfaces has not been practically investigated. There are only a limited number of works where laser treatment has been combined with formation of self-lubricating films on ceramic surfaces (Voevodin et al. 1998). The micro-reservoirs were machined by a focused UV laser beam on the surface of hard TiCN coatings (Voevodin and Zabinski 2006). The surfaces of hard coatings were then filled with MoS_2 and graphite films. Burnishing and sputtering were used to deposit solid lubricant films on the laser-textured surfaces.

Bonded molybdenum disulfide (MoS_2) lubricant films are widely used in different applications, especially in spacecraft and launch vehicles. MoS_2 coatings usually show low friction (0.02–0.04) in dry and vacuum conditions and high friction in humid environments (Bhushan and Gupta 1991). Burnished films created by a rubbing process transfer solid lubricant onto the contact surfaces. Burnishing of MoS_2 or other solid lubricants is widely applied in order to improve the tribological properties of roughened substrates. To improve the tribological properties of solid lubricant films under different environmental conditions (humidity, vacuum, high temperature, etc.), MoS_2 or WS_2 powders are mixed with other powders before burnishing.

Wear life is one of the main parameters in the analysis of the efficacy of solid lubricant films. It is determined as the number of cycles (or time) of sliding to reach a high value of the friction coefficient (failure of the film).

In recent years, laser texturing has been combined with incorporation of solid lubricant into micro-reservoirs. Solid lubricant stored in the dimples can be released to the interface and thus increase the longevity of rubbed surfaces.

Preparation of Burnished Films

Sintered alumina flat samples with grain size 1–3 μm are burnished by solid lubricant IF-WS₂ nanoparticles with an average size close to 120 nm. The solid lubricant film of IF-WS₂ and platelets of MoS₂ are shown in Fig. 1. The surfaces of alumina and steel are ground ($R_a = 0.2 \mu\text{m}$). The samples are rinsed in hexane and acetone using an ultrasonic bath and then dried before burnishing. A small portion of IF-WS₂ nanoparticles are spread evenly on the surface of a cloth. Alumina samples are rubbed for 5 min in reciprocal motion with a slight pressure. The films before and after friction as well as the surface of the substrate are characterized by Raman spectroscopy, TEM, and SEM analyses.

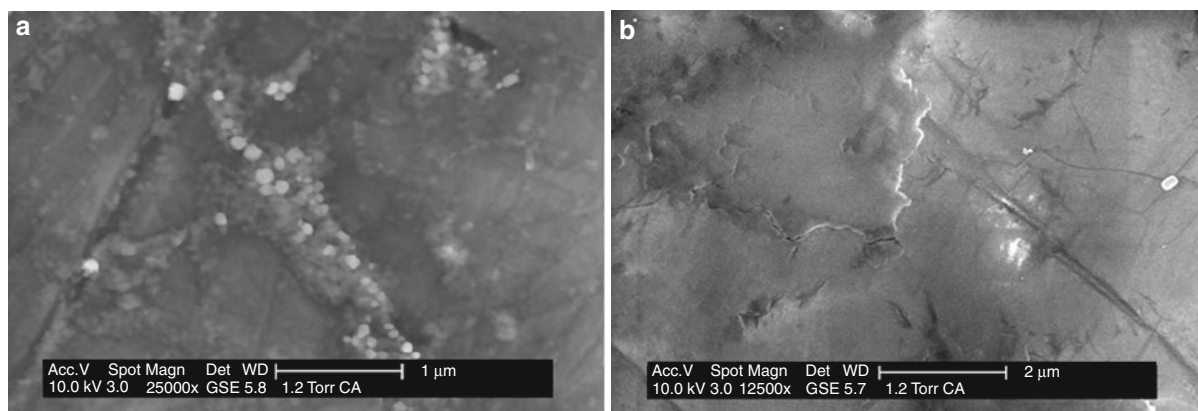
Cloth burnishing is also used to deposit a thin solid film of MoS₂ on the surfaces of textured steel disks obtained by laser surface texturing (LST) and pulsed air arc treatment (PAAT). The depth of the dimples on the steel textured surfaces is set low, about 2 μm , and the diameter of dimples is about 50 μm . The samples are lapped after LST so that the height of the bulges around the dimples is about 0.5 μm . Commercially available MoS₂ powder (<2 μm) is used to burnish steel surfaces. The thickness of solid lubricant films is about 1 μm .

Friction tests on ceramic materials are performed using a ball-on-flat device. Si₃N₄ bearing balls with a diameter of 2 mm slide against an alumina flat. Maximal contact pressure is about 1.6 GPa. Experiments have been carried out at a sliding velocity of 0.2 mm/s and a load of 0.75 N. Friction coefficient, diameter of the contact spot, and width of the wear track are evaluated during the friction test.

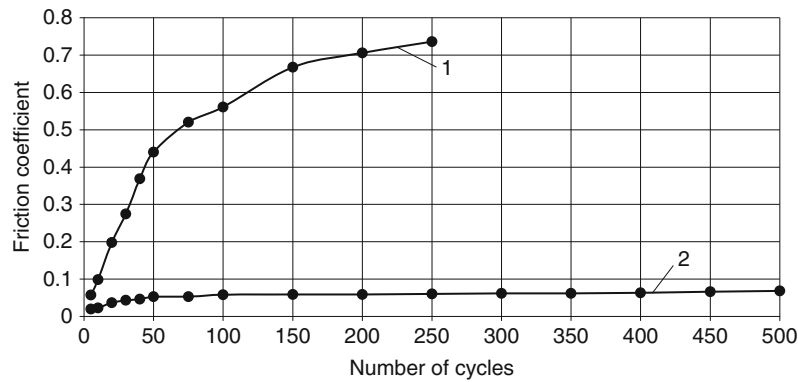
Friction tests on textured surfaces are performed using a ball-on-disk device at a sliding velocity of 0.25 m/s. Friction coefficient, diameter of the contact spot on the surface of ball, and width of the wear track on the laser-treated samples are studied. To evaluate the effect of LST density on wear life of the storage films, a load is changed by steps in accordance with the following scheme. Originally, the load is increased by steps of 15 N each minute up to a load of 90 N. After that, the load is increased by steps of 36 N each minute up to a load of 670 N. An abrupt increase in the friction force under a load of 670 N is used as a measure of the wear life of solid lubricant films. The test is stopped when the friction coefficient obtains a value close to 0.3. The number of cycles up to stopping the test indicates the wear life of the solid lubricant film.

Burnished Film on the Surface of Ceramic

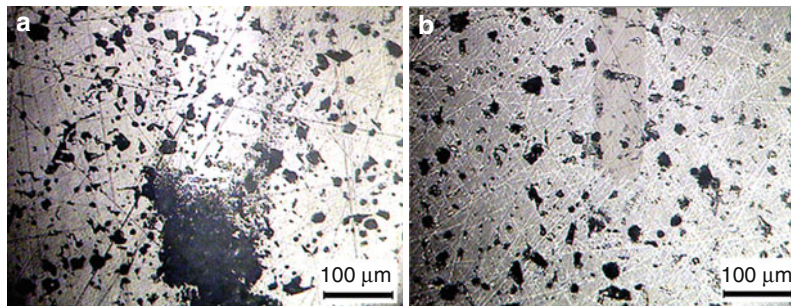
The surface of alumina burnished with solid lubricant IF-WS₂ nanoparticles is shown in Fig. 1a (Rapoport et al. 2005). The aggregates of solid lubricant nanoparticles spread on rubbed surfaces are seen in Fig. 1b. Figure 2 demonstrates the dependence of the friction coefficient on the number of cycles for the pure alumina and alumina burnished by IF-WS₂ nanoparticles. The friction coefficient increases quickly and, after friction during 250 cycles, is more than 0.7, while the friction coefficient



Solid Lubricant Films Deposited by Burnishing, Fig. 1 (a) IF-WS₂ nanoparticles in the pores of alumina after burnishing. (b) Solid lubricant film spread on the surface of alumina after friction against a Si₃N₄ ball



Solid Lubricant Films Deposited by Burnishing, Fig. 2 The variation of the friction coefficient on number of cycles for pure alumina (1) and alumina burnished with solid lubricant film (2)



Solid Lubricant Films Deposited by Burnishing, Fig. 3 The ends of wear tracks under friction of pure alumina (a) and burnished by solid lubricant film (b)

remains low ($\mu \sim 0.05$) and stable during 500 cycles. The wear tracks after friction are shown in Fig. 3.

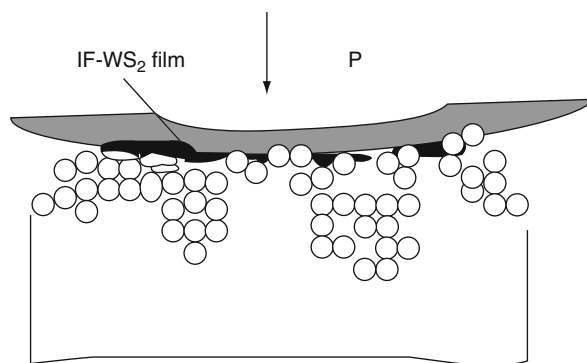
Large amounts of wear debris under friction of alumina indicate severe contact conditions leading to high friction and wear. Wear debris is practically absent on the burnished surface. Wear debris formation can be attributed to fracture of rough summits and development of cracks in the edges of pores as stress concentrators. In contrast, IF-WS₂ nanoparticles are preserved in the ranges near rough summits, fill the valleys and pores of sintered alumina, and thus limit direct contact between ceramic surfaces. A schematic illustration of ball-flat contact is shown in Fig. 4. With loading, a thin film of deformed and exfoliated nanoparticles is obtained. It is important to note that, although the external sheets of outlet layers of IF-WS₂ nanoparticles are peeled off, leading to formation of a burnished layer, many full-shape IF nanoparticles can be preserved in the valleys and pores of the contact surface and support contact over a long test period. Thus, it may be concluded that burnishing of IF-WS₂ provides excellent

tribological properties of alumina-silicon nitride pairs under high contact pressure.

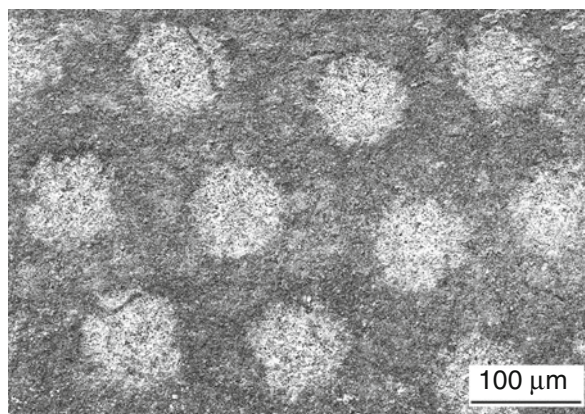
Burnished Film on Textured Surfaces

Solid lubricant film of MoS₂ platelets on the LST surface is shown in Fig. 5 (Rapoport et al. 2008).

Increasing the density of dimples usually leads to an increase in wear life. According to the test, the transition to seizure occurs after 6 min at a load of 90 N for a sample with density of 10%, while it happens under a load of 666 N for samples with densities of dimples of 40–50%. In the latter case, when a larger amount of solid lubricant powder is burnished onto a textured surface, the higher density of dimples is responsible for increased wear life of a solid lubricant film. The solid lubricant film is usually spread over the entire contact range, especially on the LST surface with a high density. The change of the friction coefficient with time for the samples with the density of dimples of 26% and 42% is shown in Fig. 6.



Solid Lubricant Films Deposited by Burnishing, Fig. 4
Schematic illustration of ball-flat contact when burnished film limits direct contact between ceramic surfaces



Solid Lubricant Films Deposited by Burnishing, Fig. 5 SEM micrograph of MoS₂ film burnished on an LST steel surface

In comparing the friction behavior of LST surfaces with dimples of 2 μm depth with reference ground surfaces, it has been found that the wear life of burnished film on ground surfaces is very low in comparison with laser-textured surfaces. The wear life of MoS₂ film on ground surfaces is less than 15 min, so the transition to seizure occurs at relatively low loads in comparison with textured samples. An EDS analysis confirmed the presence of solid lubricant film both in the dimples and in the space between dimples. The transition to seizure is associated with attrition of solid lubricant film in the space between the micro-reservoirs and the shearing of solid lubricant from the dimples. A detachment of the solid lubricant film begins on the surfaces of bulges around the dimples (Fig. 7). It is expected that seizure inception occurs

originally on the surface of bulges. Removal of solid lubricant film from around the dimples leads to increased friction and cracking of solid lubricant film in the dimples. Finally, seizure occurs in some places where the amount of solid lubricant is limited.

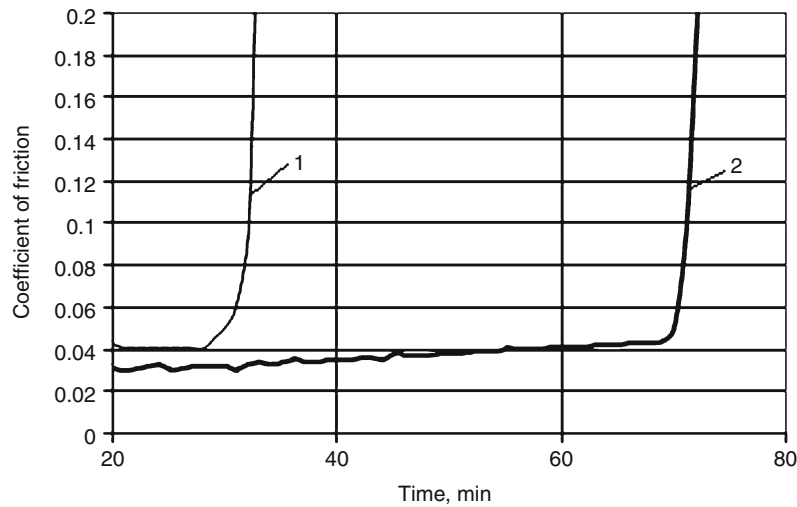
To improve adhesion between solid lubricant particles and steel substrate on the ground surface around the dimples, the efficacy of application of a sub-layer as a substrate to MoS₂ film has been studied. Some types of nanoparticles of burnished selenides and sulfides were studied as a film for the sub-layer: MoS₂, MoSe₂, CdZnSe, ZnSe₂, and PbSe (Rapoport 2009). The size of nanoparticles varied from 5 to 100 nm. A thin burnished film of CdZnSe showing the highest value of the friction coefficient was chosen as the material for a sub-layer.

To analyze the adhesion between CdZnSe and MoS₂ burnished films, a steel ball burnished with MoS₂ powder was slid slowly against the pure steel surface or the surface burnished with a thin CdZnSe film. The surfaces of ball and flat after friction during one cycle are shown in Fig. 8. The strong transfer of MoS₂ on the surface of the CdZnSe film occurred, while very small pieces of MoS₂ were observed on the steel surface. The change in the friction force during half of the first cycle is shown in Fig. 9. It can be seen that sliding between the MoS₂ film and a smooth steel surface began after a short time without any visible static friction (adhesion) (Fig. 9a). A strong adhesion between MoS₂ and CdZnSe led to static friction (Fig. 9b). Sliding began after the strong adhesion and transfer of the MoS₂ film to the surface of the CdZnSe sub-layer.

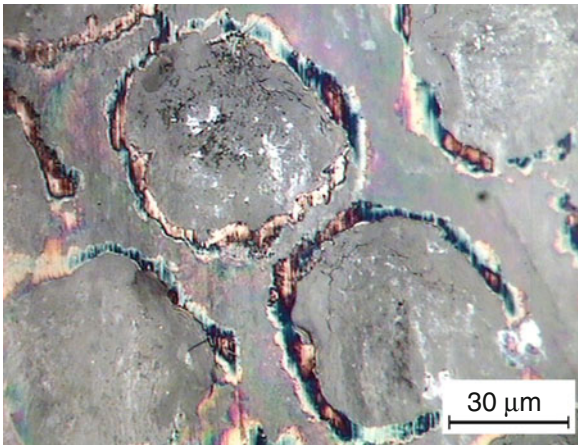
The results of these experiments indicate good adhesion between CdZnSe and MoS₂ in comparison to the friction between flat steel and the MoS₂ film burnished on the surface of the ball. The transfer of MoS₂ layers on the surface and the high value of the static friction at the beginning of the motion indicates strong adhesion between MoS₂ and CdZnSe. Strong adhesion can be associated with structural affinity between the hexagonal (wurtzite) and the hexagonal laminar structure of MoS₂.

Figure 10 presents the surfaces burnished with MoS₂ powder on the CdZnSe sub-layer. A much smoother and denser film is observed for the burnished MoS₂ film on the CdZnSe sub-layer (Fig. 10a), where all the grooves of the virgin surface are almost filled by the solid lubricant. Some grooves of the virgin ground surface are not filled completely, and some large aggregates of MoS₂ particles are observed after burnishing (Fig. 10b).

It is anticipated that high adhesion of the MoS₂ film to the CdZnSe sub-layer leads to formation of a much smoother and denser solid lubricant film.



Solid Lubricant Films Deposited by Burnishing, Fig. 6 Change of the friction coefficient with time for the samples with density of dimples of 26% (1) and 42% (2). Load, $P = 666 \text{ N}$



Solid Lubricant Films Deposited by Burnishing, Fig. 7 A detachment of solid lubricant from the bulges (around the dimples)

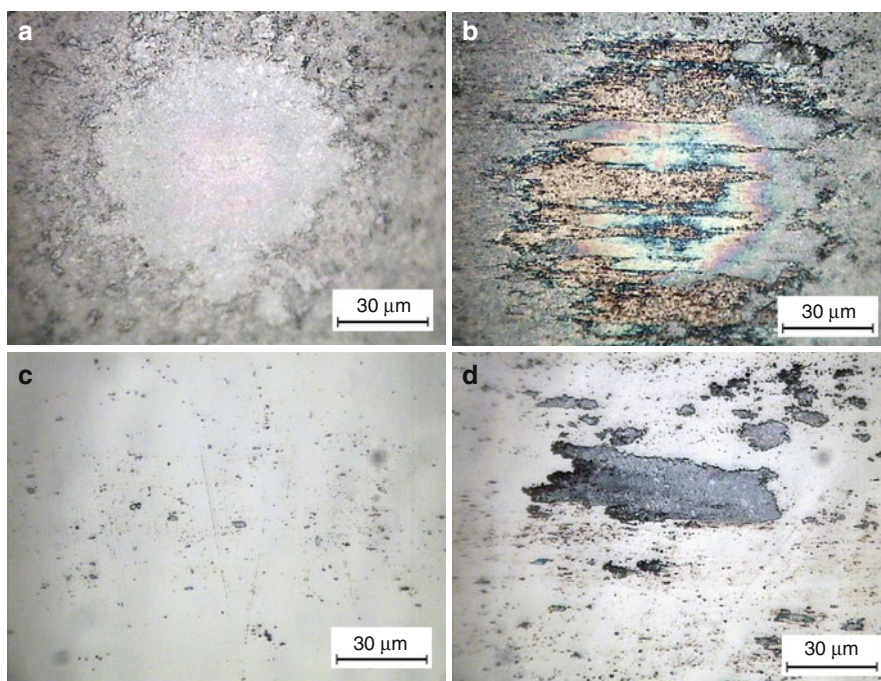
The results of the study of wear life of solid lubricant films on the rough surface without dimples are shown in Fig. 11.

The test was stopped when the friction coefficient rose to the value of 0.3. It can be seen that the wear life of the MoS_2 film on the CdZnSe sub-layer is close to 65 min, while a transition to seizure begins after about 20 min for burnished film without sub-layer.

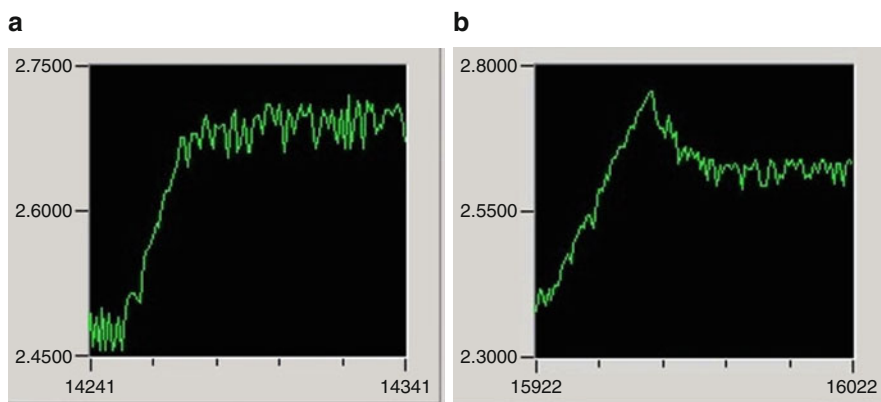
Based on an analysis of the presented results, a model of the burnishing is given in Fig. 12.

The adhesion of the MoS_2 particles to a steel surface under burnishing is mainly determined by mechanical trapping of the particles in a rough asperity profile. High surface energy of small-size of CdZnSe nanoparticles ($\sim 40 \text{ nm}$) leads to aggregation. The aggregates of nanoparticles can better fill the microgrooves of asperity surface, thus increasing the bearing ratio of the surface. MoS_2 platelets can better cover the surface of the sub-layer. The relatively low adhesion, density, and orientation of MoS_2 particles with a size of about $2\text{--}5 \mu\text{m}$ accelerates the delamination of the solid lubricant layers and thus decreases their wear life. An increase in adhesion between the sub-layer and the MoS_2 powder provides a definite orientation of lamellar MoS_2 layers and increases the density of burnished layer. Adhesion between the sub-layer and the MoS_2 powder results in the high density of the burnished film. It is expected that the high adhesion between the sub-layer and the solid lubricant film and high density of this layer limits the damage to solid lubricant film during shearing of this sheet and their transfer from one contact surface to another, thus increasing the wear life of the burnished layer.

Thus, it is seen that the formation of solid lubricant film on the textured surface increases wear life remarkably in comparison to burnished film on ground surfaces. The wear life of MoS_2 film on a CdZnSe sub-layer is more than twice that of MoS_2 film on a steel substrate.



Solid Lubricant Films Deposited by Burnishing, Fig. 8 Micrographs of the MoS_2 film on the surface of the ball rubbed with a pure steel surface: (a) the MoS_2 film after contact with the CdZnSe sub-layer, (b) the surface of flat steel sample, (c) CdZnSe sub-layer, (d) after contact with the MoS_2 film on the surface of the ball

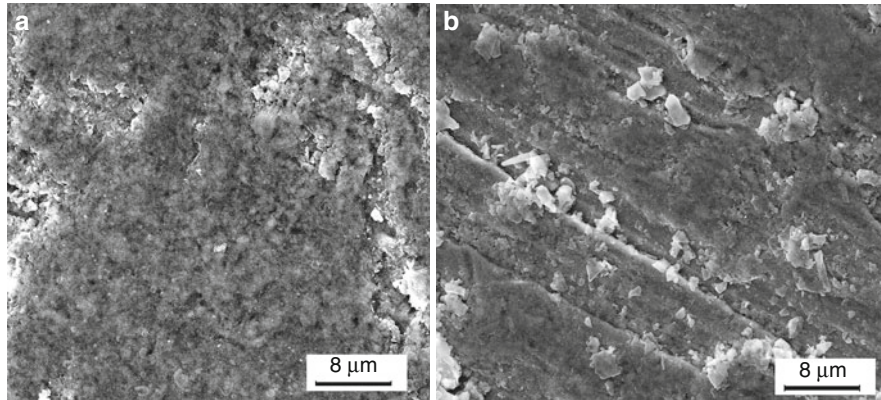


Solid Lubricant Films Deposited by Burnishing, Fig. 9 Friction force during half of the first cycle. (a) Friction between the MoS_2 film and the steel surface. (b) The MoS_2 film and the CdZnSe sub-layer

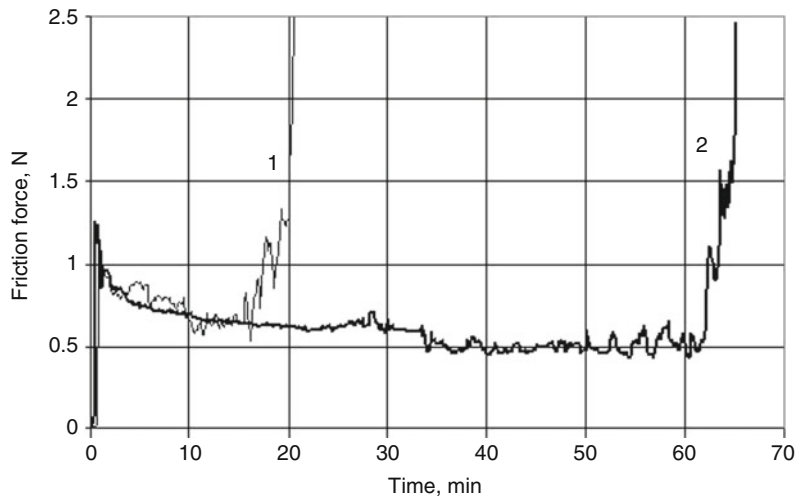
Key Applications

The main purpose of solid lubricant films is to provide easy shearing of thin lubricant films and to limit direct contact between rubbed asperities. The efficacy of

solid lubricant films is especially apparent when these films cover hard surfaces. These can include ceramics, hard coatings, or hardened steels when the plastic deformation of substrate is limited. Burnishing of solid



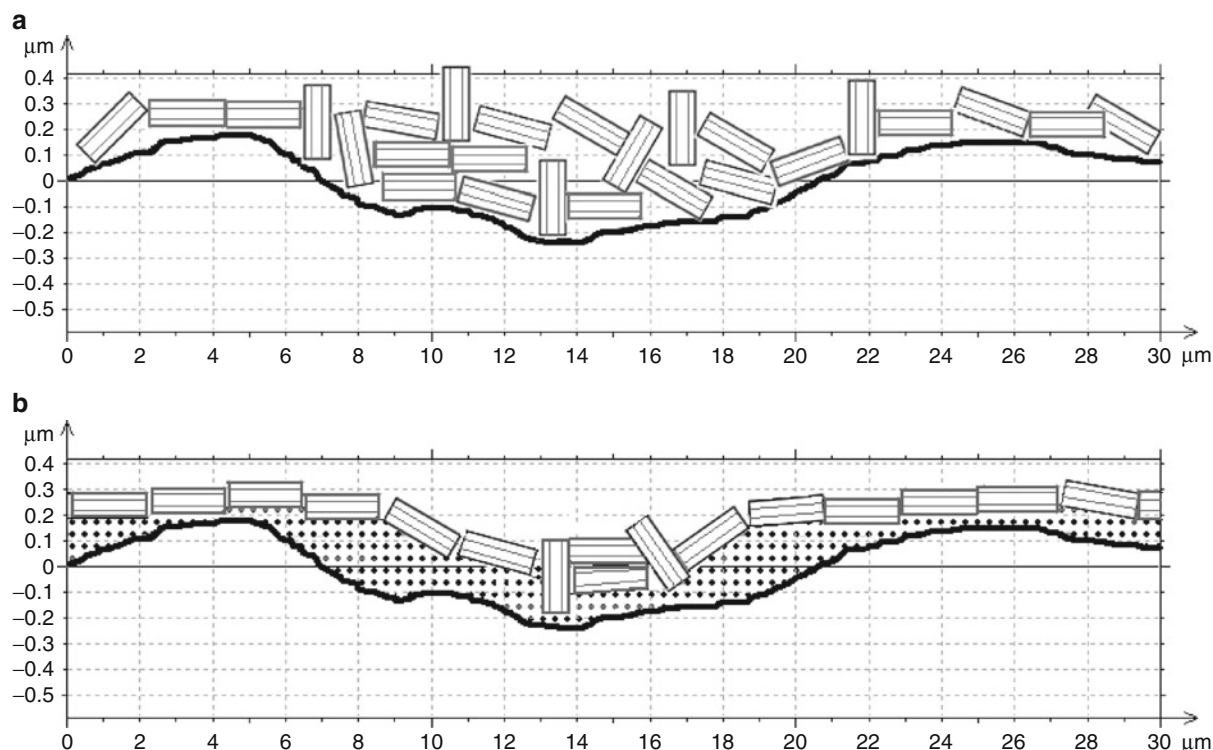
Solid Lubricant Films Deposited by Burnishing, Fig. 10 Micrograph of the surfaces burnished with (a) CdZnSe + MoS₂ film, (b) MoS₂



Solid Lubricant Films Deposited by Burnishing, Fig. 11 Dependence of the friction force versus time (1) for the MoS₂ film burnished on a steel substrate; (2) on a CdZnSe sub-layer

lubricant powder on rough or textured surfaces increases the longevity of different rubbed pairs, for example, bearings parts. Micro-reservoirs filled with solid lubricant particles provide low-friction tribological coatings for aerospace applications. This effect is based on easy supply of solid lubricant particles from the dimples into the interface under severe contact conditions. A mixture of solid lubricant powders of dichalcogenites like WS₂ or MoS₂ with different organic and inorganic binders allows creation of adaptive lubricant films for different

environmental conditions (humidity, temperature). Our experiments show that formation of the transfer film of MoS₂ or WS₂ on rubbed surfaces under friction in oil with nanoparticles provides low friction over a relatively long time (10–15 min) when the supply of lubricant into the contact interface is suddenly stopped. The formation and preservation of solid lubricant film is important under friction of contact pairs rubbed under severe contact conditions, for example, some types of high-loaded engines.



Solid Lubricant Films Deposited by Burnishing, Fig. 12 Model of the burnished film of MoS_2 (a) on a steel substrate and (b) on a CdZnSe sub-layer of nanoparticles. Virgin roughness of steel ($R_a = 0.1 \mu\text{m}$)

Cross-References

- [Bonded Solid Lubrication Coatings, Process, and Applications](#)
- [Chameleon or Smart Solid Lubricating Coatings](#)
- [Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides](#)

References

- B. Bhushan, B.K. Gupta, *Handbook of Tribology: Materials, Coatings, and Surface Treatments* (McGraw-Hill, New York, 1991)
- L. Cizaire, B. Vacher, T. Le Mogne, J.M. Martin, L. Rapoport, A. Margolin, R. Tenne, *Surf. Coat. Technol.* **160**, 282 (2002)
- I. Etsion, *Wear* **267**, 1203 (2009)
- R.L. Fusaro, *ASLE Trans.* **25**, 141 (1982)
- L. Margulis, G. Salitra, R. Tenne, M. Talianker, *Nature* **365**, 113 (1993)
- L. Rapoport, Yu Bilik, M. Homyonfer, S.R. Cohen, R. Tenne, *Nature* **387**, 791 (1997)
- L. Rapoport, Y. Feldman, M. Homyonfer, H. Cohen, J. Sloan, J.L. Hutchison, R. Tenne, *Wear* **225–229**, 975 (1999)
- L. Rapoport, M. Lvovsky, I. Lapsker, Y. Feldman, A. Zak, R. Tenne, *Adv. Eng. Mater.* **3**, 71 (2001)
- L. Rapoport, O. Nepomnyashchy, I. Lapsker, A. Verdyan, Y. Soifer, R. Popovitz-Biro, R. Tenne, *Tribol. Lett.* **19**, 143 (2005)
- L. Rapoport, A. Moshkovich, V. Perilyev, I. Lapsker, G. Halperin, Y. Itovich, I. Etsion, *Surf. Coat. Technol.* **202**, 3332 (2008)
- L. Rapoport, A. Moshkovich, V. Perilyev, A. Gedanken, Yu Koltypin, E. Sominski, G. Halperin, *Wear* **267**, 1203 (2009)
- C. Shahar, D. Zbaida, L. Rapoport, H. Cohen, T. Bendikov, J. Tannous, F. Dassenoy, R. Tenne, *Langmuir* **26**, 4409 (2010)
- R. Tenne, L. Margulis, M. Genut, G. Hodes, *Nature* **360**, 444 (1992)
- A.A. Voevodin, J.S. Zabinski, *Wear* **261**, 1285 (2006)
- A.A. Voevodin, J. Bultman, J.S. Zabinski, *Surf. Coat. Technol.* **107**, 12 (1998)

Solid Lubricant: Soft Metal

SHANHONG H. WAN

State Key Laboratory of Solid Lubrication, Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences, Lanzhou, People's Republic of China

Synonyms

[Solid lubricants based on soft metals](#)

Definition

Soft metals, a special class of material with low hardness, have been investigated as surface engineering materials used as solid lubricants in key sliding and rolling mechanical components for reducing friction and improving anti-wear ability as well as increasing equipment service lifetime. This critical review mainly addresses the tribological performance of some typical soft metals along with the operating conditions based on some highlighted characteristics of these soft metals. Some key industrial applications as well as future prospects are also discussed.

Scientific Fundamentals

Introduction

Soft metals, such as indium, silver, gold, tin, lead, and related alloys, have been extensively investigated as surface engineering materials in the form of films or coatings. These metals can provide effective isolation on the surfaces of materials, thereby reducing friction, providing a lubrication effect, preventing seizure, and creating high wear resistance for materials from room temperature to $\sim 1,100^{\circ}\text{C}$ as well as in vacuum and air atmospheres (Bowden and Tabor 1950). Thus, soft metals play a significant role in solid lubrication. The exceptional anti-wear and anti-friction properties of soft metals depend on the three characteristics of microstructure and properties of soft metals summarized in Table 1.

- The face-centered cubic (FCC) phase structure of soft metals resulting in the isotropy for the crystalline lattice essentially provides the highly viscous, fluid-like lubricating behavior.

Solid Lubricant: Soft Metal, Table 1 Basic properties of soft metals (Lansdown 2004)

Element	Melting point ($^{\circ}\text{C}$)	Crystalline form	Hardness (HB)
Au	1,063	FCC	18
Ag	960.8	FCC	25
In	156.6	FCC	3
Pb	327.3	FCC	4
Cd	321	–	20
Sn	231.9	BCC ($>16^{\circ}\text{C}$); diamond-type ($<10^{\circ}\text{C}$)	5

FCC represents face-centered cubic, and BCC represents body-centered cubic

- Low shear strength of soft metals makes the interior slip easily, resulting in the interesting self-repairing character.
- Low evaporation rate of soft metals serves in varied temperatures, from room temperature to $\sim 1,100^{\circ}\text{C}$ as well as in different ambient atmospheres.

Therefore, such high-performance soft-metal coatings are often fabricated onto sliding and rolling mechanical components usually exposed to friction, wear, and corrosion conditions, which effectively increases the service lifetime of key mechanical components and improves the reliability and longevity of the whole machine operation. This is also important in preventing the economic loss and sometimes catastrophic failure. Therefore, soft metal coatings have been widely applied in mechanical transmissions involving sliding friction and rolling friction, machine systems, and the nuclear and aerospace industries, which involve special service circumstances such as high and/or low temperatures, high speeds, high vacuum, and radioactive atmosphere.

When mechanical transmission devices and machine systems operate in extreme environments, in order to best satisfy the technological requirements in these frictional systems, the selection of the optimum form in soft metal lubricants with high performance is important. Different forms of soft metal lubricants are briefly introduced as follows:

- Coatings on harder substrates created by different processing techniques such as ion plating, magnetron sputtering, and the conventional electroplating processes, etc., improve tribological performance to meet the requirements of high temperature and highly operating pressure, exceeding the load-bearing capacities of conventional oil and grease.
- Additives in the nano-scale dispersed in liquid-based mediums such as oil, grease, and water. A tribo-chemically protective film with excellent self-lubricating performance, such as nano-copper or nano-zinc, improves anti-wear and anti-friction properties.
- Dry powder, a direct and effective application method, can improve the running-in conditions of sliding mechanical components. However, the poor adhesion of particles to the substrates is detrimental to service lifetime, especially in continuous applications.
- The introduction of soft metals into hard ceramic composites such as TiN and CrN may improve tribological performance of ceramic composite matrix by reducing the coefficient of friction and improving

anti-wear action due to their unique self-repairing behavior (Kelly et al. 2010).

Based on the above-mentioned, the poor adhesion of dry powder and the easy volatility of liquid-based lubricants fundamentally restrict their wide-scale application, especially in harsh working conditions. Therefore, soft metal coatings with excellent lubricating performance and good wear resistance are the “ideal” form for a wide range of functional applications. Furthermore, increasing requirements for adaptability of mechanical components to alternating and harsh environments, along with sustainable environmental development and the progress of nanotechnology, have been the driving forces behind development of novel alloy deposits and metal-based composite deposits to satisfy the complex service needs of current and future engineering materials.

Besides the influence of the microstructure and inherent properties of soft metals on their friction and wear properties, the tribological performance of soft metal coatings involves the following important parameters: film thickness, the specific ratio of shear stress of coating to substrate, the roughness of both substrate and the paired counterface, operating conditions (velocity, load, stroke, and ambient atmospheres, etc.), as well as the different depositing techniques (Holmberg and Matthews 2009). For example, the optimum thickness of soft metal film is vital to retard the wear of substrates by preventing plastic deformation and crack nucleation in harder substrates. So, in the following sections, the tribological performance of soft metal coatings and related alloys is reviewed on the basis of their operating parameters; some recent progress of soft metal and related alloys will also be presented in brief.

Indium Coating

Indium coatings have mainly been used in key aircraft parts to prevent them from wearing out or reacting with oxygen in air. Note that good anti-wear and anti-friction performance can be obtained at optimum thickness of indium coating fabricated onto substrates in air and vacuum environments. When indium deposits on a steel disc slide against a steel ball in air atmosphere, an even lower coefficient of friction of 0.05 presents in the initial state; however, the coefficient of friction abruptly increases up to 0.5 after 20 revolutions because of the transition of oxidation typically characterized by a shiny metallic to a grayish color, and finally the disrupting film takes place. The coefficient of friction decreased with increasing load regardless of the varying velocity within an appropriate range, however, increasing velocities greatly speed up

the onset of service failure of indium coating. For indium coating in vacuum environment, increasing sliding speed increases the coefficient of friction and wear rate (Holmberg and Matthews 2009). More recently, a “tribo-coating” of indium coating exhibited the excellent tribological performance of friction coefficient of <0.02 in sliding contacts and 0.002 in ball bearings because of good in-situ self-restoration and on-demand controllable lubrication, which is most important to improve reliability and overcome unexpected tribological troubles for the future space system (Adachi and Kato 2008).

In particular, the increasing technological requirements of high-performance frictional systems have required the development of novel multi-functional alloy coating. For Ag-In and Au-In binary alloy coatings fabricated onto steel substrates, it is interesting to note that the greater the free energy of formation of the binary alloy, the lower the friction and wear for interaction of the alloy with the iron surface, which is dominated by the tribo-chemistry interactions (Buckley 1971). Furthermore, Cu-Ni-In ternary alloy coating has been successfully fabricated in titanium alloys in aircraft turbo-machinery to improve fretting fatigue resistance and oxidation resistance at high temperatures (Chamort et al. 1988). Moreover, the addition of a small amount of indium into nitrides results in an increase in the hardness of the material as well as a decrease in the friction coefficient, along with an increase in the oxidation resistance to improve the forming efficiency.

Cadmium Coating

Cadmium coating demonstrates equal tribological performance to that of indium and silver coatings. It is remarkable that an optimum film thickness of coating onto harder substrates can effectively slow down wear loss and decrease the coefficient of friction. On the other hand, studies shows that operating circumstances have a great deal of influence on the anti-wear and anti-friction properties of cadmium coating. For cadmium coating serving in argon atmosphere, a minimum coefficient value of friction was achieved when this coating thickness was in the range of 0.1–1 μm . However, in air condition, serious degradation of cadmium coating took place and the wear process was critically dominated by extensive delamination wear caused by oxidation. Furthermore, it is important to point out that surface roughness moderately influences film service lifetime. For instance, polishing may extend the sliding distance, yet polishing also weakens the adhesion strength of deposits to substrate; thus, an optimum roughness on substrates plays an important role in improving wear resistance without sacrificing bonding

strength. Recently, stricter environmental and human health regulations have come into effect, and such cadmium coatings are now being used much less often in modern industry applications because of pollution. Other environmentally friendly metals or alloy coatings may be capable of replacing cadmium coating for wear and corrosion protection (Holmberg and Matthews 2009).

Lead Coating

Lead coatings have been widely investigated as an engineered surface material to improve lapping, decrease friction, and protect machine parts against corrosion. As for its tribological performance, it is typically characterized by a high friction coefficient of 1–2, normal presence of large-scale seizure, and tearing closely related to its easy plastic deformation. As also true for indium and silver coatings, the tribological performance of lead coating is dependent on the optimum film thickness, the counterpart materials, and metal substrate hardness, as well as surface roughness. When a 30 μm thick lead deposit on a copper substrate was sliding against a steel ball, a decreasing coefficient of less than 0.3 was obtained at elevated temperature of 327°C; however, a thinner or thicker lead coating on steel substrates possessed an increased coefficient value of friction. The different wear mechanisms were dependent on film thickness, e.g., microcutting for the thicker one caused by the mating surface asperities, and fatigue mechanism for the thinner one. Furthermore, if the counterpart is changed to electropolished copper, a 10 μm thick lead coating on copper substrate can provide a minimum value of friction coefficient. As for high surface roughness of the tribo-paired materials, a thinner lead film makes the interaction of substrate to the slider easier. In summary, the inherent factors of lead coating affect its friction and wear properties.

Besides the inherent influence of the tribo-paired material, the anti-wear and anti-friction ability of lead coating, in particular, responds sensitively to normal load and sliding velocity. If the applied load and sliding velocity are in the optimum range, a minimum value of friction coefficient is obtained. Remarkably, a longer wear lifetime of ion-plated lead coating is provided when operating at higher velocity, which is related with the lubricating effect of lead oxide forming at elevated temperatures.

Especially in space components, lead coating demonstrated interesting anti-wear and anti-friction performance in high vacuum atmosphere. A minimum value of the coefficient of friction was obtained for the film thickness within the range of 0.2–1 μm with surface

roughness less than 0.5 μm (Ra), however, the coefficient of friction increases with increasing film thickness as well as rougher surface. Moreover, there is a controversy that the coefficient of friction increases with increasing speed in vacuum. It should be noted that thicker lead films all showed an increasing coefficient of friction regardless of surface roughness, ranging from 0.05 to 0.46, meanwhile the wear lifetime of lead film was considerably extended (Holmberg and Matthews 2009).

Tribological performance in air is different from that in vacuum, which is believed to be related with the different surface geometry. On the contrary, a minimum coefficient value of friction and a decreasing wear life of this coating at highest load were found in air environment, as well as that in vacuum atmosphere. At elevated temperature, lead film operated in a vacuum showed a very low friction coefficient of 0.06 but very short service lifetime, which is believed to be because increasing load and elevated temperature make lead melting easily, resulting in lower shear strength but increasing wear. If the operation temperature exceeds the melting point of lead, it leads to earlier failure of this coating. However, even at cryogenic temperatures of 20K in vacuum, lead lubricating coating bearings display equal tribological performance as at room temperature because of excellent cold shortness (Holmberg and Matthews 2009).

Lead coating provides desirable lubricating performance within the optimum speed range in a space environment, especially for ball bearings. However, the generated debris during the operation process seriously limits the service lifetime, and high-precision application of lead film, detrimental ball-speed variations caused by lead film, and ball-to-ball friction cause a high level of torque noise. Lead coating with an optimum thickness can also upgrade the load-capacity of lubricated sliding steel contacts to some degree, typically characterized by the increasing scuffing load as well as low coefficient of friction. Furthermore, it is worth noting that the friction of lead film showed satisfactory adaptability in a wide range of sliding speeds and applied loads.

Conventional single metal coatings do not satisfy the increasing technological demands of high-performance mechanical systems, so it is important to develop multi-layered or alloy coatings for enhancing adhesion strength and extending wear life. Using other metals as the interlayer, including Mo, Ta, W, Ag, and Cu, and only a thin copper interlayer at the lead/steel interface can provide adequate improvement. On the other hand, when such metals as Mo, Ag, Au, and Pt were incorporated into lead deposits in small amounts, only copper or platinum added to lead coating effectively extended the wear

life – by about fivefold. Stricter environmental and human health regulations have placed some restrictions on a commonly used lead coating; currently, lead as the essential component surface engineering coating has been applied in exploiting other more environmentally friendly alloy coatings to improve key frictional systems in many industrial applications, e.g., binary alloys such as Pb-Ag, Pb-Sn, and Pb-Zn (Hombostel 1991).

Gold Coating

Gold coating typically represents good electrical and thermal conductivities, as well as excellent corrosion- and wear-resistance. Therefore, gold coating has been widely applied to electrical connectors and related components in electrical industry. Note that the primary requirements in these applications are electrical conductivity and endurance lifetime rather than tribological performance. However, intensive attention has been paid to gold coating, mainly in the aerospace industry, and this coating possesses excellent anti-wear and anti-friction ability in special working conditions. A thin gold coating fabricated onto steel cylinders at an optimum thickness of 0.1 μm can greatly reduce the wear loss by three orders of magnitude in sliding motion against each other, as compared to uncoated cylinders. It was surprising that gold coating was able to reduce wear loss in air and argon atmosphere, however, the coefficient of friction maintained high at value of about 0.85–0.9.

Moreover, the desired tribological performance of electroplated gold coating on steel substrates can be obtained at the optimum operation conditions, such as appropriate film thickness, optimum velocity, and a suitably applied load. As compared to uncoated steel contacts, a stable service lifetime was remarkably extended by four orders of magnitude for thicker coatings, ranging from 5 to 20 μm when the testing parameters were controlled at the velocity of 0.0132 m/s and a high applied load of 680 N. Furthermore, when a small amount of nickel is doped into electroplated gold deposits, the wear resistance and hardness of this Au-Ni composite coating can be further improved.

On the other hand, especially in electrical connectors with excellent electrical conductivity, the service stability of this coating is also very important. When gold coating is sliding against itself, the wear process undergoes three states; an initial friction coefficient of 0.3 and then 1.2 dominated by the different wear transition from the initially marginal wear, subsequent cracks at the nickel/brass interface caused by the large interfacial shear stress, and finally enlarged delamination at the interface. Further, adopting the solid-liquid composite method, the

coefficient of friction for gold coating can decrease when using polyphenyl ether as the lubricant. However, the fretting resistant performance of gold coating showed some interesting properties. The coefficient of friction maintained no great change regardless of the changing contact system in dry and lubricating conditions as well as different frequencies and amplitudes (Holmberg and Matthews 2009).

Especially for some key frictional components in extremely harsh working environments, more advanced materials are urgently needed to meet increasing technological demands without sacrificing the inherent properties of gold coating. Therefore, some Au-based or alloy coatings have been developed to improve the longevity and duration stability of high performance machines. The example of Au coating with the incorporation of MoS_2 in different amounts indicates that the anti-friction and endurance life of Au coating has been effectively regulated, and an optimum MoS_2 content in this composite coating would possess a low wear loss event at high contact pressure, which is mainly attributed to the good lubricating effect of MoS_2 . For Au-base alloy coating, alloying gold generally increases the strength and hardness but with a slight sacrifice of malleability and ductility as compared to pure gold, such as Au-Ag, Au-Cu, and Au-In, etc. Au-Ag alloy coating sliding steel in argon atmosphere shows a discontinuous increase in friction and wear, accompanied by the onset of metal transfer related with iron solubility in sliding contacts. Such Au-Cu (In) coatings have been applied in bearings, which can provide good operation conditions such as resistance to high pressure, operation at high linear speed, resistance to high temperature, as well as good service lifetime. Furthermore, electroplated Au-Cu-Pd ternary alloy coating has been described as promising in materials for electric/electronic applications for contacts subjected to sliding, fretting, or repeated insertions, so the tribological performance is critical for the functional performance in both kinds of industrial applications. When this coating is sliding against a hard countermaterial, the wear process is dominated by abrasion caused by plastic deformation, namely micro-ploughing. And it worth stressing that the higher ratio of adhesion energy to Young's modulus, the higher the resistance of Au-Cu-Pd ternary alloy coating to adhesive failure (Bozzini et al. 2003).

Silver Coating

Thin silver coating has been widely applied to engineering mechanical components subjected to sliding or rolling operating conditions due to its unique physical properties, such as its chemical inertness, good oxidation resistance,

and very high thermal conductivity as well as its easy shear to reduce the friction.

As compared with other single soft metal coatings, including lead and indium, silver coating possesses superior anti-friction capability. It is stressed that the degradation of silver coating is mostly due to abrasive wear, because hard oxide particles are formed by the rising temperature closely related with increasing friction by a typical characteristic of the coefficient of friction from initial 0.15 to 0.5. Moreover, the service life of silver film in air is constant if the velocity is relatively low, otherwise the lifetime is shorter.

Good lubricating ability of silver coating is attributed to the synergy effect of low shear strength in the sliding direction and a material transfer between the tribo-paired interfaces. For silver coating onto a ball, a high wear loss of the coated ball occurs due to micro-cutting caused by surface asperities in the initial stage; subsequently, a continuous removal of silver coating is caused by a non-uniform smearing of silver flakes around the track, and then a uniform material transfer leads to a steady wear loss, and finally the direct contact of ball to disc takes place when the silver coating completely breaks down. Therefore, during the sliding process a great number of fine abrasive particles like steel and iron oxide are produced, and play an important role in polishing; however, this greatly restricts its further applications, especially for high-precision components. Contact fatigue causes low wear loss between the original film and substrate interface.

On the other hand, the anti-wear ability and failure mechanism of silver coatings are closely dependent on the contact pressure and the sliding speeds, which have been confirmed by two-layer gradient silver coating on steel substrates against a steel ball using a ball-on-disk tester. It was found that there are three different wear regimes: (1) mild wear and elastic-plastic deformation without failure of coating; (2) moderate wear and formation of transfer layer at the contact; and (3) severe wear and no protective transfer layer. It is necessary to point out that this functionally gradient silver film composed of an 8 nm thin IBAD silver bond layer and a 2 μm thick thermally evaporated silver film remarkably extended the service lifetime, which was mainly attributed to better bond strength, as compared with single silver film by thermal evaporation and IBAD. Moreover, the mild wear regime was ascribed to transfer layers of agglomerated wear particles on the contact surfaces, and this transfer layer mainly acted as a protective layer, resulting in low friction after the initial transfer of coating material (Yang et al. 2003).

Silver coatings have been successfully fabricated by different processing techniques, including ion plated, vacuum deposited and gas deposited processes. Among them, ion-plate silver coating was the most suitable applied to certain materials' surfaces for the purpose of anti-wear and anti-friction. Because this coating shows distinct nucleation, small size of nuclei, high density, and can provide larger endurance life and lower coefficient of friction of about 0.15, as compared with other two processing techniques, it results in good adhesive strength of coating to substrates. Moreover, the operating environments considerably influenced the anti-friction and anti-wear properties of silver coating, as well as the deformation. Such hydrophilic and hydrophobic silver coatings fabricated onto the rolling components demonstrated interesting tribological performance compared with unmodified silver coating when rolling against steel balls in different humid atmospheres. There was no variation in the initial rolling process regardless of different humid atmospheres. However, some discrepancies in the coefficient of friction existed in the steady state among those samples. A feasible explanation is that adsorbed water vapors influence the wear debris agglomeration and the tendency of contacting patch formation, resulting in the discrepancy in the rolling resistance. Furthermore, a thin silver film can greatly reduce the sliding wear loss by three orders of magnitude at the optimum velocity and applied load in argon atmosphere. Surprisingly, an increasing coefficient of friction occurs for both thinner and thicker silver films, which has no great effect on the film service lifetime, while the increasing sliding speeds result in increasing coefficient of friction but decreased wear life (Holmberg and Matthews 2009).

Mechanical components in modern industry face increasing performance requirements, leading to the growing need for advanced Ag-based materials and, thus, for modern friction systems. For example, single Ag coating onto ceramic components including Al_2O_3 , Si_3N_4 , magnesia-partially stabilized zirconia, and zirconia exhibited good tribological performance in air, i.e., a remarkably lowered coefficient of friction by about 50% and a decreasing wear loss by 1–3 orders of magnitude, especially for zirconia and Al_2O_3 . Furthermore, using a thinner titanium interlayer greatly improved the adhesion strength and the effectiveness of silver coating to Al_2O_3 substrates to improve the anti-friction and anti-wear ability in air and in humid atmosphere. The addition of solid lubricants into Ag matrix such as MoS_2 and graphite can provide good anti-friction characteristics of the rubbing surface of

sliding pairs, especially for weak-current sliding pairs with high operating life (Braterskaya et al. 1991). Furthermore, silver alloy coatings are generally characterized by high mechanical and multi-functional performance as well as high electrical conductivity in modern industrial applications, especially in key aircraft parts, and are also stable at elevated temperature, such as Ag-Sn and Ag-Pd based alloys (Kwon et al. 2009). Moreover, adding precious Ag metal to hard nitride matrix (e.g., TiN, CrN, and Mo₂N) exhibited outstanding self-lubricating ability even at higher operating temperature, however, this result was also accompanied by a corresponding decrease in hardness (Kelly et al. 2010).

Tin Coating

Legislative pressures and customer demands have been the driving force behind the use of lead-free coatings in the connector industry. More specifically, a tin or tin-alloy coating can effectively reduce the friction force generated within an electrical connector, minimizing insertion and fretting wear. However, it is very susceptible to the fretting conditions, and further severe degradation may occur due to wear resulting from the accumulation of debris within the contact region, reducing the conductive contact area and causing an increasing voltage drop across the electrical connector, further resulting in connector failure. Thus, multilayered coatings and composite coatings have been exploited to increase hardness and wear resistance and to reduce friction without sacrifice of the electrical performance. Typical examples include Cu-Ni-Sn and related Sn-based coatings as well as Sn-PTFE composite. For Cu-Ni-Sn coating, using a nickel layer as the interlayer can greatly improve tribological performance of the non-noble tin-tin electrical contact, which is a favorable effect of nickel interlayer related to its capacity to modify the friction behavior under partial slip in sliding condition regime. However, copper readily oxidizes at elevated temperature, resulting in deterioration of some electrical connectors. If PTFE particles are introduced into ductile tin matrix with an optimum compositional range and within an appropriate particle size range, this composite coating has excellent electrical characteristics and good wear resistant properties (Guenin and Conn 1991; Jedrzejczyk et al. 2009).

Key Applications and Future Prospects

In engineering, soft metal lubricants have been used in almost every type of components using PVD, CVD, and conventional electroplating processes in mass production, evolving in many important industrial fields such as automotive, aircraft and other transportation industries,

electronics, consumer goods, and metals manufacturing industries, as well as chemical, food, pharmaceutical, medical, and packing industries for the purpose of reducing the coefficient of friction and extending the wear life as well as the cost and environmental protection. Soft metal coatings play a significant role in many key sliding and rolling frictional components, such as sliding bearings, rolling bearings, gears, seals, cams and tappets, pistons, cylinders, valves, injectors, plungers, rotors, pumps, and transmissions.

With developments in industry and nanotechnology, requirements for the high performance of modern engineering frictional systems become more stringent and complex, which motivates the exploitation of the highly functional performance of advanced soft metal-based coatings. Based on the above-mentioned discussion, attention has been focused on developing high-performance deposits as well as the environmentally friendly processing techniques in the future. In particular, many soft metal lubricants have been exploited by modern emerging nanotechnology, from the original single metal deposits to high-performance advanced coatings, to better satisfy high-performance requirements in modern frictional systems, using the new designing concept including nanocrystalline, functionally graded, and multilayered materials.

Cross-References

- [Cryogenic Solid Lubrication](#)
- [Gear Lubricants](#)
- [Solid Lubricants](#)
- [Thin Film Lubrication](#)

References

- K. Adachi, K. Kato, SM/SEED space exposure experiment of ball bearing lubricated by tribo-coating, in *Proceedings of International Symposium on "SM/MPAC & SEED Experiment,"* Tsukuba, 2008, pp.121–125
- F.P. Bowden, D. Tabor, *Friction and Lubrication of Solids. Part I* (Oxford University Press, Oxford, UK, 1950), pp. 321
- B. Bozzini, A. Fanigliulo, E. Lanzoni, C. Martini, Mechanical and tribological characterisation of electrodeposited Au-Cu-Pd. *Wear* **255**, 903–909 (2003)
- G.N. Braterskaya, T.A. Dontsova, E.A. Zaitsev, V.P. Smirnov, Use of silver-base electrospark coatings in sliding contact pairs. *Powder Metallurgy Met. Ceramics* **30**, 44–47 (1991)
- D.H. Buckley, *Thermochemistry of Binary Alloys and Its Efficient Upon Friction and Wear* (National Aeronautics and Space Administration, Washington, DC, 1971)
- C. Chamort, Wear problems in small displacements encountered in titanium alloy parts in aircraft turbo-machines, in *Sixth World Conference on Titanium*, France, 1988
- B.M. Guenin, G. Conn, Composite coating for electrical connectors, U.S. Patent No. 5028492, 1991

- K. Holmberg, A. Matthews, *Coating Tribology: Properties, Mechanisms, Techniques and Applications in Surface Engineering*. Tribology and Interface Engineering Series (Elsevier Science, Amsterdam/Boston/London, 2009), pp. 197–211
- C. Hombostel, *Construction materials: Types, Uses, and Applications* (Wiley–IEEE, New York, 1991), pp. 485–487
- P. Jedrzejczyk, S. Chad, S. Fouvry, P. Chalandon, Impacting of the nickel interlayer on the electrical resistance of tin-tin interface submitted to fretting loading. *Surf. Coat. Technol.* **203**, 1624–1628 (2009)
- P.J. Kelly, H. Li, P.S. Benson et al., Comparison of the tribological and antimicrobial properties of CrN/Ag, ZrN/Ag, TiN/Ag, and TiN/Cu nanocomposite coatings. *Surf. Coat. Technol.* **205**, 1606 (2010)
- J.D. Kwon, S.H. Lee, K.H. Lee, J.J. Rha et al., Silver-palladium alloy deposited by DC magnetron sputtering method as lubricant for high temperature application. *Trans. Nonferrous Met. Soc. China* **19**, 1001–1004 (2009)
- A.R. Lansdown, *Lubrication and Lubricant Selection* (ASME Press, New York, 2004), p. 171
- S.H. Yang, H. Kong, E. Yoon, D.E. Kim, A wear map of bearing steel lubricated by silver films. *Wear* **255**, 883–892 (2003)

Solid Lubricants

KAZUHISA MIYOSHI

Consultant, North Olmsted, OH, USA

Synonyms

Dry film lubricants; Dry lubricants; Solid film lubricants

Definition

A solid lubricant is a solid material that provides lubrication between two surfaces moving in relation to one another (Campbell 1972). A solid lubricant is basically any solid material, such as a thin film or a powder, that can be placed between two bearing surfaces and that will shear more easily under a given load than the bearing materials themselves (Lansdown 1996).

Scientific Fundamentals

Solid Lubrication Mechanism

The classical Bowden and Tabor model for sliding friction, in its simple form, assumes that the friction force arises from two contributing sources (Bowden and Tabor 1954). First, an adhesion force is developed at the real area of contact between the surfaces (the asperity junction). Second, a deformation force is needed to plow or cut the asperities of the harder surface through the softer. The resultant friction force is the sum of the two contributing sources: friction due to adhesion and friction due to

deformation and/or fracture. The adhesion arises from the attractive forces between the surfaces in contact. This model serves as a starting point for understanding how thin surface films can reduce friction and provide lubrication (Rabinowicz 1995). It should be realized, however, that one of the contributing sources acts to affect the other on many occasions. In other words, the two sources cannot be treated as strictly independent.

Solid lubrication is achieved by imposing a solid material or self-lubricating material having low shear strength and high wear resistance between the interacting surfaces in relative motion. Reducing the coefficient of friction requires minimizing the shear strength of the interface, the surface energy, the real area of contact, and the plowing or cutting contribution (Rabinowicz 1995; Miyoshi 2001). Reducing wear generally requires minimizing these factors while maximizing the hardness, strength, and toughness of interacting materials. Toward this end, surface design and engineering of solid lubricants can be applied to reduce the coefficient of friction and wear rate of materials.

The Most Commonly Used Solid Lubricants

Depending on the nature of the two surfaces, a wide variety of solid materials can reduce friction. However, the ability to reduce friction is only one of the important properties of solid lubricants, and most materials have serious faults that rule them out as effective lubricants. As a result, the vast majority of solid lubricant applications are met by only a few materials: graphite, molybdenum disulfide (MoS_2), polytetrafluoroethylene (PTFE), soft metals, and their bonded solid lubricants (composites). Graphite and MoS_2 have lamellar structure, and PTFE is a fluorocarbon solid consisting wholly of carbon and fluorine. The most commonly used solid lubricants and their characteristics are summarized in the following.

Graphite has a low coefficient of friction and possesses very high thermal stability. It has a hexagonal crystal structure with the intrinsic property of easy shear, although graphite relies on adsorbed moisture or water vapors to achieve low friction. Use in dry environments, particularly in vacuum, may be limited. At temperatures as low as 373 K, the amount of water vapor adsorbed may be reduced to the point that low friction cannot be maintained, so sufficient water vapor may be deliberately introduced to maintain low friction. Practical application at high temperatures is limited to a range of 773–873 K because of oxidation. When necessary, additives composed of inorganic compounds may be added to enable use at temperatures to 823 K.

MoS₂ has a low coefficient of friction both in vacuum and atmosphere, and it does not rely on adsorbed vapors or moisture. Its thermal stability in nonoxidizing environments is acceptable to 1,373 K, but in air the temperature limitation of MoS₂ may be reduced to a range of 623–673 K by oxidation. Adsorbed water vapors and oxidizing environments may actually result in a slight, but insignificant, increase in friction. MoS₂ has greater load-carrying capacity than other commonly used lubricants, such as graphite and PTFE. MoS₂ has a hexagonal crystal structure with the intrinsic property of easy shear. The lubrication performance of MoS₂ often exceeds that of graphite, and MoS₂ is effective in vacuum where graphite is not.

PTFE has a low coefficient of friction both in vacuum and atmosphere because of a lack of chemical reactivity (Buckley 1981). PTFE does not rely on adsorbed vapors or moisture. It possesses low surface energy and does not have a layered structure. The macromolecules of PTFE slip easily along each other, similar to lamellar structures. Practical application temperatures range from 173 to 523 K. PTFE does not have greater load-carrying capacity and durability than other alternatives. The low thermal conductivity of PTFE inhibits heat dissipation, which causes premature failure due to melting and limits use to low-speed sliding applications where MoS₂ is not satisfactory. PTFE shows one of the smallest coefficients of static and dynamic friction, down to 0.04. Operating temperatures are limited to about 523 K.

Soft metals must have a high degree of plasticity for easy shear and to adequately adhere to bearing surfaces in relative motion. Soft metals, such as lead, gold, silver, copper, indium, tin, and zinc, possess relatively low coefficients of friction both in vacuum and atmosphere because of their low shear strengths (Buckley 1981; Lansdown 1996; Miyoshi 2001). Also, they are good conductors of heat and electricity. All metals except gold can oxidize in air, and friction and wear behavior of metals may be affected by oxidation. The metals are extremely useful for high-temperature applications, mainly in vacuum and for rolling element applications, such as roller bearings, where sliding is minimal. Although gold and silver films have been used in spacecraft applications, gold is preferred to silver because of oxidation issues. Silver and barium films have been used successfully on lightly loaded ball bearings in high-vacuum X-ray tubes. Lead and copper powders are also used in anti-seize compounds for high-temperature use.

Bonded solid lubricants (solid lubricating composites) are combinations of two or more different solid phases mixed together in which one or more of the solid

lubricants, such as PTFE, MoS₂, graphite, or soft metals, are embedded in the major phase called the *matrix phase*. The matrix material can be metal, ceramic, or polymer. There are mainly two types of bonded solid lubricants, resin-bonded solid lubricants and inorganic bonded solid film lubricants (Campbell 1972). Resin-bonded solid lubricants are probably the most common and widely used solid lubricants types, even today. The resin-bonded solid lubricants usually consist of a lubricating solid (pigment) and a bonding agent (binder or matrix) and are generally applied in thin films to the surfaces of the metal components being lubricated. The function of the pigment is to provide the low friction and wear reduction required for the system being lubricated, and the binder serves to hold the lubricating pigment to metal surface. Although methods of attaching a lubricating solid to a bearing surface are many and vary considerably, the end result is the same, that is, a low-friction medium is deposited to reduce friction and wear between two relatively moving surfaces under essentially dry conditions. For bearing use, most polymers are strengthened by carbon or glass fiber, and many of them have PTFE or MoS₂ added to reduce friction. The inorganic bonded solid film lubricants are usually referred to as high-temperature solid lubricants. They employ ceramics or nonceramics (salt-based binders) to give greater temperature resistance than resins and usually employ lubricating solids, such as calcium fluoride, gold, silver, tellurides, and selenides, which are more thermally and oxidatively stable than graphite or MoS₂. There are exceptions, however, and a number of the ceramic and salt-based binders are used with graphite or MoS₂.

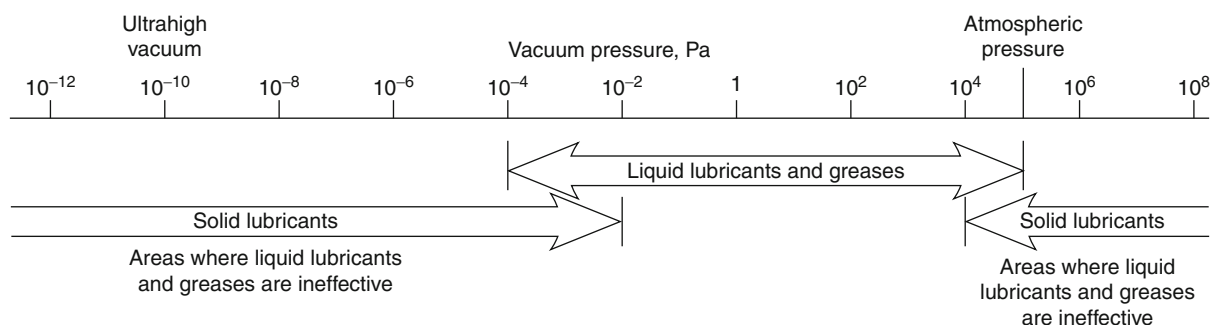
Advantages and Disadvantages of Solid Lubricants

Table 1 shows some advantages and disadvantages of solid lubricants (Campbell 1972; Miyoshi 2001). Under high loads, intermittent loading, high speeds, high temperatures, cryogenic temperatures, high vacuums, high radiation, high dust, or corrosive environments, solid lubrication may be the only feasible system. In addition, Figs. 1, 2, 3, present critical operating conditions under which fluid lubricants are ineffective or undesirable, along with the most common conditions requiring the use of solid lubricants (Kakuda 1988; Miyoshi 2001):

1. In extremely high temperature conditions, liquid lubricants can decompose or oxidize. Suitable solid lubricants can extend the operating temperatures of systems beyond 523 K while maintaining relatively low coefficients of friction.

Solid Lubricants, Table 1 Advantages and disadvantages of solid lubricants

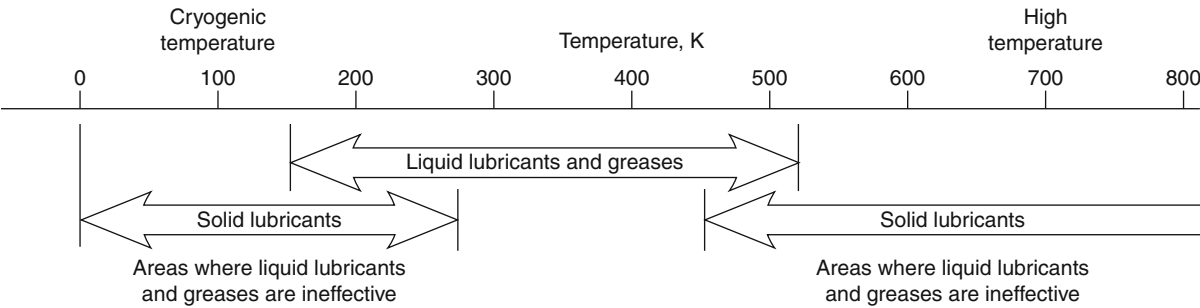
Advantages	Disadvantages
<ul style="list-style-type: none"> • Are highly stable in high-temperature, cryogenic temperature, vacuum, and high-pressure environments 	<ul style="list-style-type: none"> • Have higher coefficients of friction and wear than for hydrodynamic lubrication
<ul style="list-style-type: none"> • Have high heat dissipation with high thermally conductive lubricants, such as gold films and diamond films 	<ul style="list-style-type: none"> • Have poor heat dissipation with low thermally conductive lubricants, such as polymer-based films and ceramic-based films
<ul style="list-style-type: none"> • Have high resistance to deterioration in high-radiation environments 	
<ul style="list-style-type: none"> • Have high resistance to abrasion in high-dust environments 	<ul style="list-style-type: none"> • Have poor self-healing properties so that a broken solid film tends to shorten the useful life of the lubricant. (However, a solid film, such as a carbon nanotube film, may be readily reapplied to extend the useful life)
<ul style="list-style-type: none"> • Have high resistance to deterioration in reactive environments 	
<ul style="list-style-type: none"> • Are more effective than fluid lubricants at intermittent loading, high loads, and high speeds 	
<ul style="list-style-type: none"> • Enable equipment to be lighter and simpler because lubrication distribution systems and seals are not required 	
<ul style="list-style-type: none"> • Offer a distinct advantage in locations where access for servicing is difficult 	<ul style="list-style-type: none"> • May have undesirable color, such as with graphite and carbon nanotubes
<ul style="list-style-type: none"> • Can provide translucent or transparent coatings, such as glassy ceramic films and diamond films, where desirable 	

**Solid Lubricants, Fig. 1** Ranges of application of various lubricants in vacuum environments. (Figure has both solid and liquid lubricants)

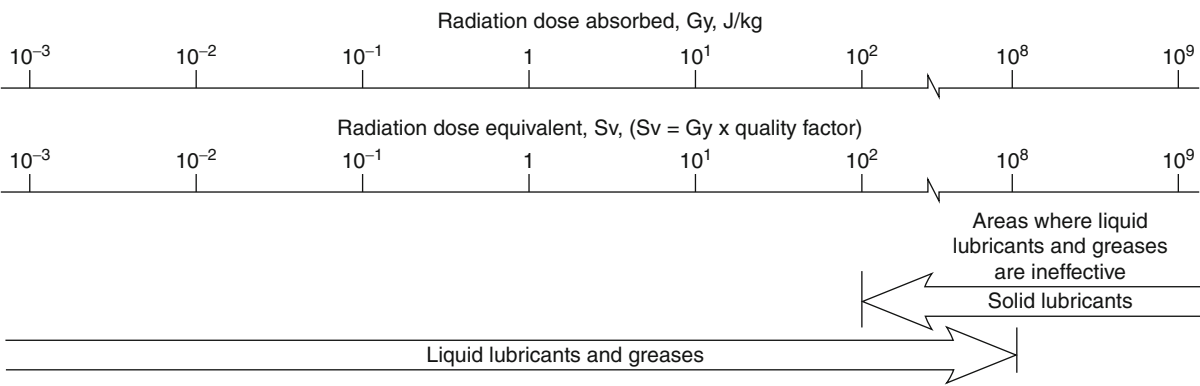
2. At cryogenic temperatures, liquid lubricants can solidify or become highly viscous and not be effective. Suitable solid lubricants can extend the operating temperatures of systems down to cryogenic temperatures.
3. In radiation environments, liquid lubricants can decompose. Suitable solid lubricants can extend the operation of systems to high-radiation beyond 10^6 rads (radiation dose absorbed of 10^4 J/kg) while maintaining relatively low coefficients of friction.
4. In high dust areas, composite solid lubricants in the hard matrix material, such as silicon carbide and

boron carbide, are useful in areas where liquid lubricants tend to pick up dust. These contaminants readily form a grinding paste, causing abrasion and damaging equipment.

5. Under intermittent loading conditions or in corrosive environments, liquid lubricants become contaminated. Changes in critical service and environmental conditions – such as loading, time, contamination, pressure, temperature, and radiation – also affect liquid lubricant efficiency. Solid lubricants can be stable in corrosive environments, high-pressure environments, and intermittent loading conditions. When



Solid Lubricants, Fig. 2 Ranges of application of various lubricants in cryogenic and high-temperature environments. (Figure has both solid and liquid lubricants)



Solid Lubricants, Fig. 3 Ranges of application of various lubricants in radiation environments. (Figure has both solid and liquid lubricants)

equipment is stored or is idle for prolonged periods, solid lubricants provide permanent, satisfactory lubrication.

6. In extreme pressure conditions (i.e., high to ultrahigh vacuum conditions – a vacuum of $\sim 10^{-2}$ Pa or higher or a gas density of $\sim 10^{-12}$ molecules/cm³ or lower at 298 K), such as space, lubricants can volatilize. In high-vacuum environments, a liquid lubricant would evaporate and contaminate the device, such as optical and electronic equipment.
7. In weight-limited equipment, such as nanotechnology-related or space-related equipment (satellites, spacecraft, and rovers), solid lubrication has the advantage of weighing substantially less than liquid lubrication. The elimination (or limited use) of liquid lubricants and their replacement by solid lubricants would reduce spacecraft weight and, therefore, have a dramatic impact on mission extent and craft maneuverability.

Key Applications

Selection and Applications for Solid Lubricants

Lubricating films have three classifications: solids, liquids, and gases (Table 2). Solid lubricants are used when liquid and gas lubricants do not meet the advanced requirements of modern technology (Table 3). Oils or greases cannot be used in many applications because of the difficulty in applying them, sealing problems, weight, or other factors, such as environmental conditions. Solid lubricants may be preferred to liquid or gas films because they reduce weight and simplify lubrication. For many applications, solid lubricants are less expensive than oil and grease lubrication systems.

Designing for Solid Lubricants

Friction contributes to energy consumption and wear contributes to short solid-lubricant lives. The important

Solid Lubricants, Table 2 Types of lubricating films

Type	Lubricating films
Solid films	Lamellar film (molybdenum disulfide (MoS ₂) and graphite)
	Polymers (polytetrafluoroethylene (PTFE), nylon, and polyethylene)
	Soft metallic film (gold, silver, lead, indium, barium, gallium, thallium, copper, tin, and zinc)
	Bonded solid lubricants or composites (resin-bonded solid lubricants and inorganic bonded solid film lubricants)
	Lamellar carbon compound film (fluorinated graphite or graphite fluoride (CF))
	Diamond and diamond-like carbon (DLC) coatings (diamond, hydrogenated amorphous carbon (a-C:H), amorphous carbon (a-C), and carbon nitride (C ₃ N ₄))
	Ceramic coatings (vanadium carbide (VC), boron carbide (B ₄ C), silicon carbide (SiC), silicon nitride (Si ₃ N ₄), titanium carbide (TiC), titanium nitride (TiN), titanium carbon nitride (TiCN), aluminum nitride (AlN), and cubic boron nitride (cBN))
	Metallic oxide film (titanium dioxide (TiO ₂), calcium fluoride (CaF ₂), silicon dioxide (SiO ₂) and related glasses, alumina (Al ₂ O ₃), lead oxide (PbO), zinc oxide (ZnO), and tin oxide (SnO))
	Nanotubes, nano-onions, and other nanoparticles (carbon allotropes, hexagonal BN (hBN), MoS ₂ , and tungsten disulfide (WS ₂))
	Nano-layered composite coatings (tungsten carbide/carbon (WC/C), MoS ₂ /C, WS ₂ /C, TiC/C, and nanodiamonds)
	Fats, soap, wax (stearic acid)
	Ceramics and cermets
Fluid films	Hydrodynamic film:
	Thick hydrodynamic film
	Elastohydrodynamic film
	Hydrostatic film
Thin films	Squeeze film
	Mixed lubricating film
Gas films	Boundary lubricating film
	Hydrodynamic film
Gas films	Hydrostatic film

parts of designing for durable, high-performance solid lubricants are to understand, decide, and take action in some or all of the following:

1. How are the solid lubricants attached to the substrate?
2. What are their strengths and surface energies?
3. How do they break down?
4. How do they self-heal?
5. How can we extend their lifetimes?
6. Can they be reapplied to the surface at service areas?
7. What are their performance benefits? Can they provide some or all of the following?

- Abrasion and wear resistance
- Hard surface
- High impact strength
- Remarkably low surface energy
- Increased thermal conductivity and thermal transfer
- High nonstick (release) properties
- Lowest friction (energy consumption) attainable
- Permanent dry lubrication to prevent galling
- Erosion protection
- Radiation protection
- Nontoxicity
- Chemical protection
- Nonwetting properties
- Precision conformance over complex geometry
- Excellent corrosion resistance
- Adequate temperature range

Desirable Characteristics of Solid Lubricants

Many of the characteristics of solid lubricants are actually surface properties. For example, friction, adhesion, bonding, abrasion, wear, erosion, oxidation, corrosion, fatigue, and cracking are all affected by surface properties (Campbell 1972; Buckley 1981; Lancaster 1984; Miyoshi 2001). Designers can enhance performance, that is, lower surface energy, adhesion, and friction, and increase resistance to abrasion, wear, erosion, oxidation, corrosion, and cracking, as well as improve compatibility with the operating environments. Further, the friction and wear properties of solid lubricants are system properties. This means that performance and behavior depend on the lubricant, the bearing materials, the operating conditions, and the system.

Areas Where Solid Lubricants Are Applicable

The coefficient of friction and lifetime of any lubricant generally vary with the environment; and lubricants have very different characteristics under different conditions. It is essential, therefore, to select the right lubrication technique and lubricant for each mechanical and tribological application.

The technology of solid lubrication has advanced rapidly in the past four decades, primarily in response to the

Solid Lubricants, Table 3 Application of solid lubricants

(a) Areas where fluid lubricants are undesirable		
Requirement		Applications
Resist abrasion in dirt-laden environments		Automobiles, off-road vehicles and equipment, construction equipment, textile equipment, agricultural and mining equipment, buildings, bridges, industrial facilities, dental implants, aircraft, space vehicles (rovers), lunar base and equipment, and Martian base and equipment
Avoid contaminating product or environment		Microscopes and cameras, spectrometers, medical and dental equipment, artificial implants, food-processing machines, semiconductor manufacturing equipment, optical equipment, metalworking equipment, surface-mounted assembly and equipment, hard disks and tape recorders, textile equipment, paper-processing machines, business machines, automobiles, space telescopes, equipment in lunar base, and equipment in Martian base
Maintain servicing or lubrication in inaccessible or hard-to-access areas		Artificial implants, buildings, bridges, industrial facilities, nuclear reactors, consumer durables, semiconductor manufacturing equipment, aerospace mechanisms, aircraft, space vehicles, and satellites
Provide prolonged storage or stationary service		Aircraft equipment, railway equipment, missile components, nuclear reactors, heavy plants, buildings, bridges, furnaces, space telescope mounts, space antenna mounts
(b) Areas where fluid lubricants are ineffective		
Environment		Applications
High vacuum	Room temperature or cryogenic temperatures	Vacuum products, analytical equipment, coating equipment, space mechanisms, satellites, space telescope mounts, space platforms, and space antennas
	Clean room	Biomedical equipment, analytical tools, coating equipment, and semiconductor manufacturing equipment
	High temperature	X-ray tubes, X-ray equipment, furnaces, space nuclear reactors, space vehicles, space mechanisms, and lunar bases and equipment
High temperatures	Air atmosphere	Furnaces, metalworking equipment, compressors, turbines, and nuclear reactors
	Molten metals (sodium, zinc, etc.)	Nuclear reactors and molten metal plating equipment
Cryogenic temperatures		Turbopumps, liquid nitrogen pumps, butane pumps, freon pumps, liquid natural gas pumps, liquid propane pumps, refrigeration plants, lunar and Martian bases, space mechanisms, satellites, space vehicles, space propulsion systems, space antennas, space-telescope mounts, space platforms
Radiation (gamma rays, fast neutrons, X-rays, beta rays, etc.)		Nuclear reactors and plants, X-ray equipment, space mechanisms, satellites, space vehicles, space platforms, space antennas, lunar base, and Martian base
Corrosive gases (chlorine, etc.)		Semiconductor manufacturing equipment and space crafts' maneuvering systems
High pressures or loads		Metalworking equipment, bridge supports, plant supports, and building supports
Fretting wear and corrosion (general)		Aircraft engines, turbines, landing gear, automobiles, buildings, bridges, industrial facilities, space antennas, and space platforms

needs of the aerospace and automobile industries. Solid lubricants are used where the containment of liquids is a problem and when liquid lubricants do not meet the advanced requirements. Under high vacuum (such as in space), high temperatures, cryogenic temperatures,

radiation, dust, clean environments, or corrosive environments, and combinations thereof, solid lubrication may be the only feasible system. The materials designed for solid lubrication must not only display desirable coefficients of friction (0.001–0.3) but must maintain good durability in

different environments, such as high vacuum, water, the atmosphere, cryogenic temperatures, high temperatures, or dust. Therefore, the successful use of materials as solid lubricants requires understanding their material and tribological properties and knowing which solid lubricant formulation is best for a chosen application. Issues such as substrate surface pretreatment, materials compatibility, the mating counterpart material, and potential debris generation must be taken into account during the design and application of a lubricated device or of moving mechanical assemblies.

Cross-References

- [Asperities](#)
- [Bonded Solid Lubrication Coatings, Process and Applications](#)
- [Bonding at Surfaces/Interfaces](#)
- [Diamond-Like Carbon Coatings](#)
- [Engine Lubricants](#)
- [Solid Lubricants for Space Mechanisms](#)
- [Solid Lubricants, Ceramic-Based Self-Lubricating Materials](#)
- [Solid Lubricants, Graphene](#)
- [Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides](#)
- [Solid Lubricants, Polymer-Based Self-Lubricating Materials](#)
- [Solid Lubrication in Fretting](#)
- [Solid-Like Lubricating Films, Self-Assembled Films](#)

References

- F.P. Bowden, D. Tabor, *The Friction and Lubrication of Solids – Part 1* (Clarendon, Oxford, UK, 1954)
- D.H. Buckley, *Surface Effects in Adhesion, Friction, Wear, and Lubrication* (Elsevier, Amsterdam, 1981)
- M.E. Campbell, *Solid Lubricants: A Survey* (National Aeronautics and Space Administration, Washington, DC, 1972). NASA SP-5059
- K. Kakuda (ed.), *NSK Technical Journal*, 648 (Nippon Seiko, Tokyo, 1988)
- J.K. Lancaster, *Solid Lubricants*, (CRC Handbook of Lubrication, Vol. II, E. R. Booser, ed., CRC Press, Boca Raton, FL, 1984)
- A.R. Lansdown, *Lubrication and Lubricant Selection – A Practical Guide* (Mechanical Engineering Publications, London/Bury St Edmunds, 1996)
- K. Miyoshi, *Solid Lubrication Fundamentals and Applications* (Marcel Dekker, New York, 2001)
- E. Rabinowicz, *Friction and Wear of Materials*, 2nd edn. (Wiley, New York, 1995)

Solid Lubricants and Applications

- [Bonded Solid Lubrication Coatings, Process, and Applications](#)

Solid Lubricants Based on MoS_x

- [MoS_x Coatings by Closed-Field Magnetron Sputtering](#)

Solid Lubricants Based on Soft Metals

- [Solid Lubricant: Soft Metal](#)

Solid Lubricants for Gas Bearings

- [Gas Bearing Materials](#)

Solid Lubricants for Rolling Bearings

- [Rolling Bearing Lubricants](#)

Solid Lubricants for Space Mechanisms

XIAOJUN SUN

State Key Laboratory of Solid Lubrication, Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences, Lanzhou, Gansu, People's Republic of China

Synonyms

[Dry lubricants for space mechanisms](#); [Oil-less lubricants for space mechanisms](#)

Definition

“Solid lubricants for space mechanisms” refers to solid-state materials (particles, thin films, coatings, and bulk materials) that can be used as lubricants to reduce friction and wear of contacting and coupling surfaces in relative motion in a space environments.

Scientific Fundamentals

Requirements of Solid Lubricants for Space Mechanisms

“Solid lubricants for space mechanisms” is a branch of solid lubricants that guarantees reliable operation of moving components in various providing stable the friction and low wear rate. Effects of space environments – including high-vacuum, elevated-temperature, cryogenic temperature, solar and/or cosmos radiation (such as high-energy particles, X-rays, ultraviolet), sand-dust, and chemical corrosive environments (such as atomic oxygen in low earth orbit) – on reliability of solid lubricants must be taken into account in space mechanisms. Thus, solid lubricants for space mechanisms simultaneously feature low friction coefficient, stable operation state, good wear resistance, and excellent space environmental adaptability. The significance of solid lubricants for space mechanism can be clarified just from Jost’s statement: “. . . even a small tribological failure can clearly lead to catastrophic results. . .” (H. P. Jost 1990).

Early application of solid lubricants aimed to resolve tribological problems (friction, wear, and lubricant) under space operating conditions, where as conventional liquid lubricants were liable to fail under the same conditions. With the progress of space exploration (such as long-term space stations and lunar rovers), demand for lubricants for space mechanisms has grown rapidly. A challenge is to match the lubricant with the key function and operating environment for each mission. Generally, space mechanisms that require lubrication including solar array drivers, momentum, reaction, and filter wheels, tracking antennas, slip rings, scanning devices, sensors, rover wheels, robotic arms, antenna arrays, gearboxes/harmonic drives, and actuators among others. Each mechanism has unique hardware and requirements based on operating environments, thus requiring unique lubrication.

Different from liquids, solid lubricants are considered the most appropriate choice for space applications. Most of the solid lubricants are applied as thin films (less than one micrometer in thickness) and thick coatings (thickness is from 1 to over 10 μm). The most commonly used solid lubricating coatings for space mechanisms include lamellar solids, soft metals, polymers, and other low shear strength and/or super-hard inorganic lubricating materials, such as DLCs and nano-structured and multi-layered thin films. Lamellar solid lubricants include transition metal dichalcogenides, such as MoS_2 and WS_2 , etc. Soft metals include Ag, In, Au, and Pb, etc. Polymers include polyimides (PI) and polytetrafluoroethylene (PTFE), etc. Sputtering and ion-plating are the preferred fabrication

methods for these lubricants. Another common method is bonded coating, in which solid lubricant particles are mixed with an organic binder and form a coating to the surface by spraying or dipping. Self-lubricating polymers and polymer-based composite are also utilized. These materials are mostly used as a retainer of rolling bearing or as a bushing. The most successfully applied solid lubricants for space mechanisms are various forms of sputtered MoS_2 based composite films and ion-plated soft metal films. The advantages of solid lubricants for space mechanisms over liquid lubricants are summarized as follows (W. R. Jones 2000).

Advantages of Solid Lubricants for Space Mechanisms

- More stable operation state than liquid lubricants while applied under elevated temperature, cryogenic temperature, ultra-high vacuum, high-radiation, and corrosive environments.
- Maintains more lasting adequate lubrication under intermittent loading, wide range of loads and speeds (especially low speed), and rerunning after a long rest.
- High resistance to abrasive wear in sand-dust environments due to successful development of super-hard lubricating films.
- Complex sealing systems for liquid lubricants are not needed for solid lubrication. So, replacement of liquid lubricants by solid lubricants reduces weight of spacecraft and benefits extension of functions and maneuverability of a mission.
- Provides reliable operation under high radiation, maintaining a relative stable friction coefficient and high wear resistance.

Disadvantages of Using Liquid Lubricants in Space

- Liquid lubricants would evaporate under ultra-high vacuum and contaminate devices such as optical and electrical units.
- Most liquid lubricants would decompose or be oxidized at elevated-temperature. Proper solid lubricants can extend the operating temperature with relatively low friction coefficients.
- Liquid lubricants would be frozen, thereby becoming ineffective.
- Liquid lubricants would decompose under solar and/or cosmos irradiation when exposed to space environments.
- Liquid lubricants tend to pick up dust, forming a grinding paste, causing abrasive wear to damage equipment under sand-dust conditions.

- Liquid lubricants become unstable even ineffective or under intermittent operating conditions.

Design Proposals for Solid Lubricants for Space Mechanisms

Based on future requirements of solid lubricants for space environment applications, several design proposals are summarized below (K. Miyoshi 2007):

- For high-accuracy moving components that need stable and precise lubrication, low friction coefficient solid lubricating films should be selected, such as modified sputtered MoS_2 -based composite films, ion-plated soft metal films, PTFE, DLCs, diamond films, multi-layered nanocomposite films ($\text{MoS}_2/\text{WS}_2/\text{C}$ and MoS_2/WS_2), and functionally graded, multilayered inorganic films (TiC_x/C).
- For wide-range temperature cycling environments, an interlayer between a substrate and a solid lubricating film should be created by controlling chemical composite and microstructure, minimizing mismatches of thermal expansion coefficient, lattice parameter, and differences in mechanical properties, strengthening chemical attraction and adhesion (e.g., a Ti or Cr interlayer on a metal substrate, a Si interlayer on a ceramic substrate, or a Zn interlayer on a polymer substrate).
- For high speeds, high loads, and other applications where heat dissipation is required, high thermal conductive solid lubricating films should be constructed, such as carbon nanotubes, nanocrystalline diamond films, metal doped diamond and DLC films, and soft metals (Au, Ag, Cu, Pb) and their alloys films.
- For a sand-dust environments, such as a lunar rover would encounter, hard solid lubricating films should be applied, such as ceramic films, including BN, BC, VC, AlN, CN, TiO_2 , SiC, Si_3N_4 , and SiO_2 , nano- and micro-crystalline diamond, and DLC films.
- For advanced controlling mechanisms, self-adaptive hard and tough solid lubricating films should be employed. These films include multiphase composite films (WC/C , $\text{WS}_2/\text{WC}/\text{DLC}$, and TiC_x/C), multilayered composite films (MoS_2/DLC , WS_2/DLC , MoS_2/WS_2 and TiC_x/C), and functionally graded films ($\text{Ti-TiC}_x\text{-DLC}$).

Key Applications

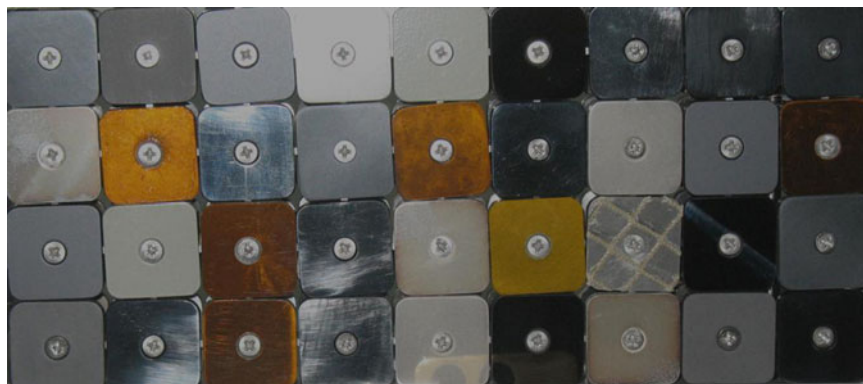
Overview

Historically, solid lubricants for space applications can be traced to the early 1960s when the demands of space

mechanism moved lubricants from liquids to solids in order to adapt to space environments. Various routes for fabrication and application of solid lubrication systems were developed. These techniques did contribute to successful. Thus, most work in this area ceased in the early 1970s. Renewed interests in lubricants for space mechanisms grew during the late 1980s and early 1990s with the emerging demands of new space missions as well as the rapid developments in advanced material science. These included novel solid lubricating films and coatings with cryogenic friction coefficient and high wear resistance, light-weight materials for moving components of spacecraft mechanisms, lubricating systems for long-term spacecraft, and retainers for bearings running in ultra-low temperature. The coming requirements of solid lubricants include stable, reliable, and long-term operation in space environments. For example, the expected endurance life of a future long-term space-station as currently under development is more than 15 years. National Aeronautics and Space Administration's (NASA) objectives for deep space exploration are human exploration of and permanent human presence on the Moon and Mars (K. Miyoshi 2007). These are critically dependent on the reliable operation of many moving mechanisms, and need advanced solid lubricant selection and fruitful efforts in applications where liquid lubricants are ineffective and inappropriate. Numerous moving components will be directly exposed to space environments during space docking and planet exploration. The effects of space environments on the tribological properties of such lubricants may pose potential threats to the overall performance of spacecraft. Thus, the environmental effects of space must be regarded when designing and selecting the required long-term space solid lubricants. Advanced solid lubricants are required to meet these operating requirements. Sand-dust wind storms on the surface of the deep space planets provide additional challenges to explorers, such as the Lunar rovers and the Martian rovers. In low earth orbits, the effects of atomic oxygen and ultraviolet irradiation should not be ignored. To explore the low earth orbit environment effects on performance degradation and failure mechanism of solid lubricants, ten solid lubricant samples (see Fig. 1) of eleven materials of three different types of solid lubricants were selected for an in-orbit, outside-cabin exposure experiment on China's Shenzhou VII manned spacecraft.

Space Mechanisms Requiring Solid Lubrication

Almost all space mechanisms have moving components require lubrication. The space mechanisms usually to require solid lubrication are listed as follows (W. R. Jones 2005).



Solid Lubricants for Space Mechanisms, Fig. 1 Section of the solid lubricants samples for in-orbit, outside-cabin space exposure experiment on China's Shenzhou-VII spacecraft (Photo provided by the State Key Laboratory of Solid Lubrication, Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences)

Electrical contact ring assemblies (ECRAs) are a typical example with unique lubrication requirements. Excessive electrical noise, usually due to surface contamination, is the most common failure mechanism in ECRAs. Thus, low speed operation and electrical conductivity are the two key factors that affect lubricant selection. Proper selection of lubricants and the electrical properties of degradation products are significant for reliable ECRA operation.

Gyroscopes, which are used to measure changes in orientation, operate at high speed with high accuracy. Waves in the bearing reaction torque, noise, and excess heat generation can cause a null position loss in the gyroscope, making the bearings a vital gyroscope component. The proper lubricants for a gyroscope should provide a high level of wear protection, and minimal friction, and they should also have low evaporation rate.

Momentum and reaction wheels have similar requirements of lubricant selection. The momentum wheels typically operate at a speed of 3,000 and 10,000 r/min. Therefore, lubricants are usually become more serious due to creep and degradation subjected to higher operating temperatures and stresses. Reaction wheels operate at low speeds. To support bearings with good lubrication, the lubricants must have good boundary lubrication characteristics. Control moment gyroscopes (CMGs) combine the aspects of gyroscopes and moment wheels to provide spacecraft attitude control. Therefore, both groups must be considered when selecting a CMGs lubricant.

Sensor bearings are used by many spacecrafts that contain rotating or dithering components for support. Correct lubricant selection is important to guarantee that the

sensor bearings are operating in the regular state to complete mission life and adapt environmental requirements.

Actuators and gearboxes are not in continuous operation for long terms, however, high stresses may be produced within long rest periods. The mechanisms cannot be re-lubricated between missions and also may be stored on ground for a long periods. Because examination and re-lubrication is often an expensive, complicated, and time consuming process, understanding the reaction between the selected lubricants and the actuator components during storage and rest time is a significant issue for ensuring the correct operation state of the actuators and gearboxes.

Planet rovers have many components that need lubrication and have unique lubrication requirements. During space travel, the lubricant is subject to low pressure and controlled temperature of the spacecraft, but once deployed, it will be exposed to extreme planet surface environments, including ultra-high vacuum, wide temperature ranges, various gaseous atmospheres, irradiation, and other environmental conditions, such as sand-dust and solid contaminants. Rover mechanisms include robotic arms to deploy instruments and manipulate the environment, mast assemblies to hold cameras and viewing devices, solar arrays to provide power, antennas and communication equipment masks, and a mobility system consisting of wheels, legs, and other moving components. In addition to the rover, the associated landing craft also has many lubricated mechanisms.

One-time operated mechanisms include satellite solar arrays or antenna deployments. Not all mechanisms require long-term lubrication, but even such one-time

operated mechanisms require critical lubricant selection. If these mechanisms fail to deploy, functions of the spacecraft will be partially or completely lost. Failure to fully understand the system, its dynamics, build-up, testing, and final operating environment effects can lead to disastrous results, as evidenced by the high gain antenna failure on the Galileo spacecraft (K. Miyoshi 1999). These mechanisms have unique lubrication demands because they only operate once, are low-speed applications, may have long rest periods before moving, and may be exposed to space environments. Generally, solid lubricants are the best choice for such mechanisms due to their low friction coefficient and ability to reside in the contact section.

Other mechanisms used in space also require lubrication. Some examples of such include solar array drives (SADs), which rotate solar arrays of spacecraft, ball, roller, and acme screw drives, and many types of gears and transmission assemblies, such as harmonic drives.

Most Applied Solid Lubricants for Space Mechanisms

Numerous solid lubricants have been developed for space mechanisms in the past decades. The most commonly applied ones in space mechanisms are as follows (W. R. Jones 2005):

Sputtered molybdenum disulfide (MoS_2) based films are widely used as lubricants in space mechanisms due to their ultra-low friction coefficients and long durability for rolling and sliding tribo-components in ultra-high vacuum. The distinct characteristic of MoS_2 is its highly anisotropic crystal layer structure, which consists of a layer of molybdenum atoms arranged in a hexagonal array. Each molybdenum atom is surrounded by equal distance of six sulfur atoms placed at the corners of a triangular prism. The distance between the layers of molybdenum and sulfur atoms is 0.154 nm, which is smaller than the distance between the adjacent sulfur atoms layers, which is 0.308 nm. Thus, the inter-lamellar layer attractions between the adjacent lamella are weak and consist basically of van der Waals force. However, the chemical bonds between molybdenum and sulfur atoms within the lamellae are much stronger covalent ones. Thus, macroscopic MoS_2 crystals easily shear along the van der Waals space between the lamella. The weak inter-lamellar bonding contributes to the low shear strength during sliding, which is reflected in the low friction coefficient (T. Spalvins 1982). Environmental factors obviously affect the friction coefficient and endurance life of MoS_2 films. In ultra-high vacuums, films that are one micrometer or less in thickness exhibit ultra-low friction coefficient (less than 0.01) (C. Donnet et al. 1996). Under normal vacuum,

the friction coefficients of the films may range from 0.01 to 0.04 and exhibit exceptionally low wear rates and long endurance lives. It is well known, however, that most components of spacecrafts/space mechanisms have to be subjected tests in air, especially in moist air, before launching. So the resistance to oxidation of the MoS_2 -sputtered film in moist air needs to studies further more. As (D. Yu et al. 1997) reported, after storage in moist air for 15 days, the substrate (440C bearing steel) surface which MoS_2 -sputtered corroded badly with lots of corroded spots on the disc surface, which are about 0.30–0.40 nm in diameter and 0.002–0.013 mm deep in the substrate, microcracks in the neighboring zones and some peeled parts of the MoS_2 -sputtered film. On the contrary, the substrate with MoS_2 - LaF_3 cosputtered on it were not corroded at all, looks like the same as that of the as-deposited samples, even after being stored for two months in the moist air with the same humidity ($\sim 100\%$ RH). Some metal elements, including Ti, Ni, Au, Pb, and Sb, were added to enhance the endurance life of MoS_2 based films. An adhesion interlayer and gradient layer(s) can be used to increase the wear life of most films (T. Spalvins 1982).

Ion-plated soft metallic films of Ag, Au, or Pb is solid lubricants selected for precision spacecraft mechanisms, where wear debris formation is critical and high reliability requirements must be satisfied (Spalvins and Buzek 1981). Ion-plated soft metallic films have three improvements over ordinary vapor-deposited films: increased endurance life, lower friction coefficient, and avoidance of catastrophic failure. The increased endurance life is attributed to superior adherence, lower friction coefficient (due to the dense, cohesive, small crystalline size), optimum film thickness, and gradual increase in the friction coefficient, which attributed to the formation of the graded interface in the film, after the film has been worn off.

Diamond-like carbon films (DLCs) are made of sp^2 - and sp^3 -hybridized carbon atoms and may contain some hydrogen ranging from less than 1% to about 50%, which affects not only their structures but also their properties. The great variety of DLC structures and compositions leads to a wide range of friction coefficients in various conditions. The friction coefficient and wear rate can be adjusted through preparation methods, incorporated alloying elements, and a variety of multilayer structures. Some DLCs exhibit friction coefficients below 0.01 in vacuum and show promise for future space applications.

Polymers have been successfully used in many applications. Many polymers can be used in vacuum and at cryogenic temperatures due to their low friction coefficients, low densities, excellent corrosion resistance,

and machinability. Commonly used polymers are introduced as follows.

- Polyimide (PI)

Polyimide (PI) is widely used in spacecraft mechanisms because of its favorable tribological performance at high gearing pressures and sliding speeds, excellent mechanical properties, and high thermal stability at temperatures up to 315°C. PI possesses some extraordinary characteristics such as excellent mechanical and insulating properties, good thermal stability and chemical inertness, high wear resistance, and radiation resistance (S. Bahadur and V.K. Polineni 2001). Certain additives (such as graphite, MoS₂, fluoride graphite, PTFE fiber, and so on) can improve the mechanical strength and the lubrication performance of PI. The incorporation of carbon fiber and solid lubricants in PI can greatly decrease the friction coefficient and wear rate in sliding against stainless steel. The combination of micro SiO₂ and carbon fibers and graphite can improve the tribological properties of the PI composites. The fillers of nano-Si₃N₄ and short carbon fibers and graphite can also improve the tribological properties of the PI-based composites.

- Polytetrafluoroethylene (PTFE)

PTFE is a commonly used polymer solid lubricant that provides a dry sliding friction coefficient lower than 0.2 sliding against stainless steel (Gregory Sawyer et al. 2003). The low friction coefficient of PTFE has been attributed to its low adhesion to the mating surface, the low shear stress needed to overcome adhesion at the interface, and to its ability to transfer on bare surfaces resulting in PTFE versus PTFE contact. PTFE can also be applied as thin sintered or resin-bonded films on metal and ceramic substrates. They have stable tribological behavior in vacuum. PTFE fibers and woven fabrics containing cotton or glass fibers are commonly used in bearings, gaskets, and seals applications.

- Polyamide-imide (PAI)

PAI is a kind of thermoplastic resin, and has very good high-temperature properties (approaching those of polyimides), good wear and radiation resistance, inherently low flammability and smoke emission, and the highest strength of any unreinforced thermoplastic. Applications include parts for internal combustion and jet engines, bearings and thrust washers, and mechanical, electrical, and electronic components.

- Ultra-high molecular weight polyethylene (UHMWPE)

UHMWPE exhibits excellent impact and abrasion resistance, which makes it useful for bearings, gears, bushings, and other sliding components. UHMWPE is widely used for low precision gears in space mechanisms. It has the highest sliding abrasion resistance and highest notch impact strength of any commercial polymers (Abdul Samad and Sinha 2010). In its bulk form, UHMWPE is highly wear resistant compared to many other polymers, such as polyetheretherketone (PEEK), polyethylene (PE), polystyrene (PS), etc. The outstanding characteristic of UHMWPE is that they can be operated from −269°C to 90°C, and even higher for short time. Since it does not liquefy at its softening point of 138–142°C, it retains excellent dimensional stability at temperatures up to 200°C.

- Polyoxymethylene (POM)

As a type of engineering plastic, POM exhibits good fatigue resistance and creep resistance, high impact strength and elastic modulus, good self-lubricant property, and wear resistance. Friction of POM with polydimethylsiloxane (PDMS) additive can be further reduced by the admixture of 5 wt% polytetrafluoroethylene (PTFE) additive (Laursen et al. 2009). POM filled with Cu exhibited increased friction coefficients and decreased wear weight loss compared with unfilled POM. POM can be widely applied as the rotation materials of mechanical and electromechanical fabricates electronic parts, automobiles, precision instruments, etc.

- Filled polymer composite

This is a modified polymeric lubricant where fibers (glass or carbon fibers) are dispersed into engineering polymers to enhance their mechanical and tribological performances. The fibers are typically 5–10 μm in diameter and can be continuous, milled, or chopped. Fiber size and orientation affect wear resistance and mechanical properties. Commonly specified fiber reinforcement materials are glass, aromatic polyamide (Kevlar), carbon, polyester, cotton, asbestos, graphite, and nylon. Carbon or glass-filled acetal and fiber-reinforced and filled PTFE (Rulon) are commonly specified for low-precision gears in space craft mechanisms. PTFE/chopped glass/MoS₂ composite is used as a retainer material for bearings. Solid particulate fillers, such as MoS₂, carbon, PTFE, PI, Ag, Cu, SiO₂, TiO₂, and Al₂O₃, are usually doped into polymers to improve their tribological properties.

- Polymer-based composite coatings

These are bonded to a metal substrate to give them the necessary stiffness to perform a useful bearing function. PTFE- and PI-based self-lubricating coatings

are widely used as bearing retainers for rolling and/or sliding surfaces in space mechanisms (M. J. Todd 1982). Vacuum pin-on-disk tests showed that the friction coefficients of ten different polymer-based lubricating coatings applied to 440 C steel substrates are less than 0.05, meanwhile their wear rate orders of magnitude are less than those in moist air (Fusaro 1988). PTFE films reinforced with glass fibers or bronze mesh are all suitable for use in vacuum, and are used for rod-end bearings and other applications in vacuum. This type of bearing finds widespread use in high loads with small movements.

Fabrication Methods of Solid Lubricants for Space Mechanisms

Many methods, such as plasma-based deposition, spray-bonded coatings, and chemical deposition, have been developed to successfully obtain solid lubricants for space mechanisms. Most of them are based on vacuum technologies, such as ion plating, sputtering, pulse laser deposition (PLD), ion beam assisted deposition (IBAD), or plasma enhanced chemical vapor deposition (PECVD). These methods result in strong adhesion between solid lubricant films and substrate materials. Recently, these advanced methods were combined with smart surface engineering practices, such as micro patterning or texturing, to achieve much improved tribological properties in applications. Ion plating and sputtering are the most basic forms of physical vapor deposition (PVD).

Ion plating is used for depositing soft metal films and achieving strong adhesion between substrate materials and films. In this technique, atoms of soft metal lubricants are thermally evaporated into argon plasma where they become ionized. The ions are then accelerated towards a negatively biased substrate where film growth occurs. The substrate bias also ensures that, during its growth, the film is continuously bombarded by argon ions and these help to reduce the contamination and loosely bound atoms in the growing film. Thus, ion plated metal films have high density, low porosity, high purity, and strong adhesion to substrates.

Sputtering is created by ion-bombardment of the source materials (targets). Direct current (DC), radio frequency (RF) and mid-frequency magnetron sputtering are three general sputtering modes for solid lubricants. Higher deposition rates are obtained by using magnetron sources. Sputtering is commonly employed for the deposition of MoS₂ and other metal dichalcogenide-based lubricating films.

Burnished films are obtained by rubbing dry lubricant powders onto a surface to be lubricated. It is the simplest

method of applying solid lubricants in the form of thin films. Burnishing can be carried out by hand, using a burnishing cloth, but it is difficult to reproducibly obtain films of the required thickness with this technique. More refined methods have been devised in which the lubricant is applied in more a controlled manner by mechanical means.

Bonded coatings are kinds of lubricants in which lubricant powders (usually MoS₂, PTFE, or PI) are attached to the substrates by binder materials. In general, the binding agent and lubricant powder are suspended in a solvent and the resulting dispersion is applied to surfaces by spraying, painting or dipping. Resin-bonded lubricants require curing before application. Coatings utilizing cellulose and acrylic resins are air-cured, while thermosetting resins need curing at high temperature. Resins in the latter category include epoxies, phenolics, polyimides, alkyds, silicones, and polyphenyl sulphide. Coating thicknesses are typically 25 μm. These coatings are improper for precision components since precise control over the coating thickness is difficult to achieve.

Plasma-sprayed is a type of coating technology in which lubricant particles are heated rapidly in a hot gaseous medium and projected at high velocity onto a surface to produce a coating. Several materials can be deposited simultaneously, and this capability has been exploited to produce multiphase coatings composed of a hard, wear-resistant matrix (e.g., Al₂O₃) with solid lubricants (e.g., MoS₂ and Ag). In space mechanism applications, this method has been used primarily for the development of high temperature coatings for use in re-entry vehicles.

Cross-References

- ▶ [Cryogenic Solid Lubrication](#)
- ▶ [Diamond-Like Carbon Coatings](#)
- ▶ [Gear Lubricants](#)
- ▶ [MoS_x Coatings by Closed-Field Magnetron Sputtering](#)
- ▶ [PVD: Ion Plating](#)
- ▶ [Solid Lubricants](#)
- ▶ [Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides](#)
- ▶ [Solid Lubricants, Polymer-Based Self-Lubricating Materials](#)
- ▶ [Thin Film Lubrication](#)

References

- M. Abdul Samad, Sujeet K. Sinha, Nanocomposite UHMWPE-CNT polymer coatings for boundary lubrication on aluminium substrates. *Tribol. Lett.* **38**, 301–311 (2010)
- S. Bahadur, V.K. Polineni, Tribological studies of glass fabric-reinforced polyamide composites filled with CuO and PTFE. *Wear* **200**, 95–104 (1996)

- C. Donnet et al., Super-low friction of MoS₂ coatings in various environments. *Tribol. Int.* **29**(2), 123–128 (1996)
- R.L. Fusaro, Evaluation of several polymer materials for use as solid lubricants in space. *Tribol. Trans.* **31**(2), 174–181 (1988)
- W. Gregory Sawyer et al., A study on the friction and wear behavior of PTFE filled with alumina nanoparticles. *Wear* **254**, 573–580 (2003)
- W.R. Jones, Lubrication for space applications, NASA/CR-2005-213424
- W.R. Jones, M.J. Jansen, Space tribology, NASA/TM-2000-209924
- H.P. Jost, Tribology – origin and future. *Wear* **136**, 1–17 (1990)
- J.L. Laursen, I.M. Sivebaek et al., Influence of tribological additives on friction and impact performance of injection moulded polyacetal. *Wear* **267**, 2294–2302 (2009)
- K. Miyoshi, Aerospace mechanisms and tribology technology-case study. *Tribol. Int.* **32**, 673–685 (1999)
- K. Miyoshi, Solid lubricants and coatings for extreme environments: state-of-the-art survey. NASA/TM-2007-214668
- T. Spalvins, Morphological and frictional behaviours of sputtered MoS₂ films. *Thin Solid Films* **96**(1), 17–24 (1982)
- T. Spalvins, B. Buzek, Frictional and morphological characteristics of ion-plated soft metallic films. *Thin Solid Films* **84**(3), 267–272 (1981)
- M.J. Todd, Solid lubrication of ball bearings for spacecraft mechanisms. *Tribol. Int.* **15**(6), 331–337 (1982)
- D. Yu et al., Variation of properties of the MoS₂-LaF₃ cosputtered and MoS₂-sputtered films after storage in moist air. *Thin Solid Films* **293**, 1–5 (1997)

Solid Lubricants, Ceramic-Based Self-lubricating Materials

JINJUN LU, JIAN SHANG

State Key Laboratory of Solid Lubrication,
Lanzhou Institute of Chemical Physics, Chinese Academy
of Sciences, Lanzhou, People's Republic of China

Synonyms

Ceramic-based/ceramic matrix self-lubricating coatings;
Ceramic-based/ceramic matrix self-lubricating
composites

Definition

Self-lubricating, an adjective, means “not requiring external application of lubrication to parts that experience friction because the lubricant is self-contained.” Ceramic-based self-lubricating materials are composites, and the most important characteristic is self-lubricating behavior. They are normally composed of two basic components, ceramic matrix and solid lubricants. Advanced structural ceramics (alumina ceramics, zirconia ceramics, silicon nitride ceramics, and silicon carbide ceramics) are good candidates for the matrix because they usually have low density and possess good properties, e.g., high hardness, high compressive strength, retention of mechanical

properties at elevated temperatures, and high resistance to chemical corrosion (Gangopadhyay et al. 1994). The above-mentioned properties enable these ceramics and ceramic-based self-lubricating materials to be applicable at high temperatures and in other harsh environments. Metal binders as well as the stabilizers are sometimes necessary and useful for constructing the ceramic matrix. Solid lubricants are an important component of ceramic-based self-lubricating materials and improve the tribological properties. Solid lubricants include precious metals, soft oxides, chemically stable fluorides, and a combination of various solid lubricants. Some carbides containing titanium, e.g., TiC and TiCN, are not solid lubricants, however, they can produce lubricious oxides on the frictional surface during sliding in an air or oxidizing atmosphere. In a broad sense, ceramic-based materials containing TiC or TiCN should be included in the category discussed here.

Scientific Fundamentals

Structural ceramics have high potential for application in tribological fields because of the particular properties mentioned above. However, it is reported that the coefficient of friction between ceramics is generally high under unlubricated conditions (Sloney and Dellacorte 1994), and it is not effective to use liquid lubricants when the ceramics are used under high temperature or vacuum environments. The high friction coefficients of these materials limit their uses as tribological components. To take advantage of the beneficial properties of the advanced structural ceramics, friction coefficient must be reduced. A strategy to improve the tribological properties of advanced ceramics is to fabricate ceramic-based self-lubricating materials. There are two or more benefits from the self-lubricating concept. The first is that the solid lubricants in the ceramic-based self-lubricating materials are reservoirs for a successive supply of the lubricating film on the frictional surface. The second is that material design can be simple because no supply of solid lubricants is needed.

Three aspects should be considered in fabricating ceramic-based self-lubricating materials: (1) material selection for the matrix and solid lubricant; (2) preparation, microstructure, and mechanical properties of ceramic-based self-lubricating materials; and (3) self-lubricating mechanism.

Material Selection for the Matrix and Solid Lubricant

Material selection of the matrix is based on two guidelines: (1) the working condition of the ceramic-based

self-lubricating materials and (2) physical, chemical, and mechanical properties of ceramics. For example, high-temperature ceramics with good strength and hardness at elevated temperatures must be selected for a high-temperature application. Another example is that ZrO_2 is not preferred as a matrix material for a working condition of 200–300°C with water vapor, because the aging of ZrO_2 in water vapor will prompt the initiation and propagation of cracks and thereby greatly degrade the mechanical property of the material.

The solid lubricants used in ceramic-based self-lubricating materials are usually soft metals (e.g., Au, Ag) or inorganic fluorides (e.g., LiF, CaF_2 , BaF_2). The CaF_2 – BaF_2 eutectics are the ones most frequently used as solid lubricants. Since a single solid lubricant provides lubrication in a limited range of temperatures, the combination of two or more solid lubricants proves to be effective in a wider range of temperatures. An example is the combination of Ag and CaF_2 – BaF_2 eutectic. The former is considered to be a good solid lubricant below 500°C, while the latter is effective at temperatures above 500°C. PbO is a good solid lubricant if its toxicity is not a problem. The Magneli phases (TiO_{2-x}) having low shear strength are also good solid lubricants.

The physical and chemical compatibilities between the matrix and solid lubricant are very important issues to be considered. Physical incompatibilities, e.g., mismatch of coefficient of thermal expansion or a large difference in melting point, make it difficult to prepare a material with enough mechanical strength. Interfacial reaction between the ceramic matrix and solid lubricant that yields brittle products at the interface or even leads to a total loss of solid lubricant is one of the chemical incompatibilities and should be avoided.

In short, material selection for the ceramic matrix and solid lubricant must be conducted from the viewpoints of both materials science and tribology.

Preparation, Microstructure, and Mechanical Properties of Ceramic-Based Self-lubricating Materials

Ceramic-based self-lubricating materials can be prepared in the form of either a bulk material or a composite coating depending on the requirement of the practical application. The bulk materials are typically fabricated by powder metallurgy. The coating materials are the same as bulk materials in ingredient and can be fabricated by many coating technologies, e.g., low-pressure plasma spraying (LPPS), high-velocity oxygen fuel (HVOF) spraying, ion-implantation, and laser surface cladding. The microstructures of ceramic-based self-lubricating

materials depend on the fabricating method. For example, PS304 (by plasma spraying) and PM304 (by hot pressing) have the same chemical composition and yet show quite different microstructures (Ding et al. 2007). As a consequence, the mechanical properties of the two materials are not identical.

Bulk Materials

A series of ceramic-based self-lubricating materials was prepared by drilling small holes in the ceramics and filling them with solid lubricants (graphite and the hexagonal boron nitride). In the past, in order to fabricate composites with uniformly distributed solid lubricant in the ceramic matrix, preparation was done by mixing a castable ceramics powder with lubricants. One of the limitations of castable composites is their low strength, which results in high wear rate (Gangopadhyay et al. 1994).

Today, powder metallurgy (pressureless sintering, hot pressing, and hot isostatic pressing) has been widely adopted to fabricate advanced materials. It is also feasible to use this technology to fabricate ceramic-based self-lubricating materials. Compared with hot pressing, hot isostatic pressing can provide mechanical strength but is an expensive technique. Spark plasma sintering (SPS) is a newly developed technology for material preparation. The SPS process, which is also called pulse electric current sintering, is a method used to produce dense alloy and ceramic powder compacts much more easily at a much shorter sintering time than that of hot pressing and hot isostatic pressing. The sintering is carried out by flowing pulse current through a graphite mold in which sample powder is set. The Joule heat generated in the mold and sample powder promotes the sintering of each specimen.

The ceramic-based self-lubricating materials prepared by powder metallurgy and SPS have a dual- or multi-phase microstructure. Because the mechanical strength of the solid lubricant is generally poor, the addition of solid lubricant leads to a reduced mechanical strength of ceramics. The additional amount of solid lubricant should be controlled to balance mechanical strength and tribological property. The advantage of ceramics containing TiC and TiCN is their high mechanical strength.

Coating Materials

The ceramic-based self-lubricating materials that are coatings are often prepared by spraying technology. The powders for plasma spraying require a special process and are quite different from those of powder metallurgy. The quality control of the coating by plasma spraying is much more complex than that of powder metallurgy.

In other words, there are many variables (arc current, gas content and flow rate, heating temperature, powder feeding rate, and irradiation time) to take into account. Considering the complicated chemical composition of ceramic-based self-lubricating materials, the powders are difficult to prepare and parameters for processing are not easy to control. The plasma-sprayed coatings generally have characteristics of certain porosity, a layer microstructure, and poor adhesive strength to the substrate. Like ceramic-based self-lubricating materials in the form of bulk materials, the machining and surface finish of ceramic-based self-lubricating coating materials should be handled carefully with special processes and tools. Surface finishing is crucial to the self-lubricating property. A good surface finishing is half the key to success.

Ion implantation of chlorine and halogen ions to titanium-based ceramics is useful for creating a self-lubricating film on cutting tools (Akhadejdamrong et al. 2003). Compressive stress can be induced on the surface and subsurface of the cutting tools, and is key to a prolonged wear lifetime. The drawback of ion implantation is the low thickness of the modified layer.

Laser surface cladding is a rapid solidification process with a cooling rate up to 10^{3-5} K/s due to its controlled heat input, small and thin layer melt pool, and heat conduction to the bulk substrate. The microstructure of the deposited layer is usually very fine and results in superior metallurgical properties. Laser cladding has demonstrated its capability to locally tailor the substrate surface to designed macro/micro structures with designed properties while maintaining the toughness and strength of the bulk substrates. However, it is a complex process and difficult to control.

Compared with powder metallurgy art, coating technologies (spraying, cladding, etc.) require special and advanced quality control. The internal stress of the coating and adhesive strength between coating and substrate are key problems.

Self-lubrication Mechanism

Lubricating Layer

The friction and wear of ceramic-based self-lubricating materials during dry sliding depend on the formation and failure of a lubricating layer on their frictional surface and/or a transfer layer on the counterpart's surface. Molybdenum disulfides and graphite are good solid lubricants and can provide a lubricating layer at the tribo-interface. Several oxides have been added to 3Y-TZP

(3 mol% yttria-doped tetragonal zirconia) as solid lubricants, but only copper oxide has an effect of reducing friction by forming a CuO-riched layer on the frictional surface (Pasaribu et al. 2003). The chemical composition, microstructure, and mechanical properties of the lubricating layer are still less known. The influence of counterpart materials on the formation of the lubricating layer should be clarified.

Lubricious Oxide

Tribochemically, any surface layer, such as oxides formed on the surface, surface contaminant, and even absorbed gas, can be a lubricant. Thus, on a hard substrate, a formed lubricating soft oxide layer, with low shear strength, results in considerable wear reduction and sometimes a decrease of friction. In the self-formation of lubricious oxides, a dry lubrication mechanism replacing liquids can be seen. This mechanism loses its effect because the hardness of ceramics decreases with increasing temperature. The low friction coefficient and wear rate can be attributed to special oxides, but researchers could not detect them structurally in the wear scars. These tribologically effective oxide layers are only few tenths up to hundreds of nanometers thick, and are therefore difficult to analyze.

Self-lubrication in dry forming and machining is based on the in situ formation of titanium base intermediate lubricous oxide tribofilm. In (Gardos et al. 1990), Magneli-phase oxide with Ti_nO_{2n-1} is identified as a lubricous oxide. Reduction of both the friction coefficient and wear volume in the wear test for Si_3N_4 -TiN, SiC-TiC, and (Ti,Mo)(C,N)-Ni is explained by in situ formation of Magneli-phase (Skopp and Woydt 1995). Many experimental studies report that the Magneli-phase oxide layer could work as a lubricious tribofilm to reduce the wear and friction. Since these oxides have a potential of shear deformation in elasto-plasticity, these in situ formed oxides are thought to work as a tribofilm to improve significantly the tribological performance.

The mechanisms of ceramic-based self-lubricating materials are complex. Some explanations for the self-lubricating effects mentioned are that the soft second phase in the ceramic matrix is squeezed out and smears on the frictional surface under the normal and frictional force during sliding (Pasaribu et al. 2003). It has been found that experimental conditions such as load, sliding velocity, and temperature have an important effect on the formation of the transfer layer (Pasaribu et al. 2005). In addition, the material of the counter body and humidity

also have an effect on tribological behavior (Tuchinskiy et al. 2000). For the formation of a thin soft layer, the transferred material should remain on the worn surface. However, the formed thin soft layer will also be removed by wear. The thickness of the layer is controlled by the balance between material transfer rate and the wear rate (Higgs III and Worniyoh 2008). For example, it has been (Gangopadhyay et al. 1994) demonstrated that the friction coefficient of Si_3N_4 sliding against steel can be reduced from 0.43 to 0.20 by drilling small holes in the surface and filling the holes with intercalated graphite. The reduction of the friction is due to the formation of a transferred film containing graphite, iron oxide, and silicates. However, the coefficient of friction of Al_2O_3 sliding against steel is not reduced with the same approach. There is a small amount of the transferred film on the worn surface and the content of graphite is very low. The difference in behavior between Si_3N_4 and Al_2O_3 appears to be related to the restricted supply of graphite at the contact area.

Key Applications

At present, the ceramic-based self-lubricating materials are applied mainly in working conditions at high temperature and high speed.

Turbines and Engines

In recent years, applications or operations involving severe temperature, high pressure, high vacuum, and chemical reactivity environments have promoted the development of ceramic-based self-lubricating materials. High operating temperatures in advanced power-generating systems, such as turbomachinery, gas turbines, and diesel engines, have imposed severe limitations on the currently available solid lubricants, particularly from the point of view of chemical stability at very high temperatures or in cases where inaccessible bearing mechanisms are required to operate reliably after long periods of inactivity (Ouyang et al. 2008).

Ceramic-based self-lubricating materials have been used in a wide range of high-temperature tribological applications. Zirconia-based ceramics are among the promising materials for tribo-engineering applications. Good wear resistance combined with high bending strength and fracture toughness is obtained. ZrO_2 -based ceramic coatings or composites, incorporating CaF_2 , BaF_2 , Cr_2O_3 and BaCrO_4 , have shown high potential for application in engines and turbines (Blau et al. 1999), as well as Al_2O_3 -based ceramics. These composites exhibit lower friction and wear characteristics at elevated temperature

and have adequate thermophysical properties, stable thermochemistry, and low shear strength properties.

Hot Hinge Joint

Woydt assessed several candidate lubrication mechanisms for tribo-systems operating at high temperature, and proposed promising substrate or coating systems for hot hinge joints of reusable vehicles (Woydt et al. 1997). The development of hot hinge joints based on this assessment has been reported, but poor wear resistance is noted as a problem. The development of hot hinge joints based on C/SiC substrate and appropriate coatings seems to be possible from the tribological point of view.

Cutting Tools

In dry machining, there will be more friction and adhesion between the tool and the work pieces. This will result in increased tool wear and, hence, reduction in tool life. In high-speed dry machining, the maximum cutting temperature of the insert involved can reach more than $1,000^\circ\text{C}$. Conversely, the limitation cutting speed is a function of the cutting tools used. It is thought that a reduction in all these problems could be achieved by using advanced cutting tool materials to reduce the heat generation by lowering the friction coefficients. CaF_2 is a well-known and widely used solid lubricant. It has physical (it prevents adhesion), chemical (it enables tribo-chemical reactions), and microstructural (it has a lamellar structure with low shear strength) influences on the tribological contact of working surfaces. The mechanism behind its effective lubricating performance is understood to be due to easy shearing along the basal plane of the hexagonal crystalline structures. Also, it is a useful addition in the production of self-lubricating ceramic composites, and is employed in various anti-wear applications.

The friction coefficient at the tool-chip interface in dry cutting of hardened steel and cast iron with $\text{Al}_2\text{O}_3/\text{TiC}/\text{CaF}_2$ ceramic tool is reduced compared with that of $\text{Al}_2\text{O}_3/\text{TiC}$ tool without CaF_2 solid lubricant (Deng et al. 2006).

Bearings

Solid lubrication has been considered for silicon nitride bearings at elevated temperatures, where impregnated graphite cages reduce the wear of silicon nitride balls by an order of magnitude.

A new ternary structural ceramics (MAX phase, so-called because they possess a $\text{M}_{n+1}\text{AX}_n$ chemistry, where n is 1, 2, or 3, M is an early transition metal element, A is

an A-group element, and X is C or N) with a layered nature suggests they may have excellent promise as solid lubricant materials. First-generation MAX phase-based composites shafts have been successfully tested against Ni-based superalloy at 50,000 rpm from room temperature to 550°C during thermal cycling in a foil bearing rig (Gupta et al. 2007). This study further demonstrates the potential of MAX phases and their composites in different tribological applications.

Cross-References

- [High-Temperature Solid Lubricating Materials](#)
- [Solid Lubricants](#)

References

- T. Akhadejdamrong, T. Aizawa, M. Yoshitake, A. Mitsuo, Feasibility study of self-lubrication by chlorine implantation. *Nucl. Instrum.: Methods Phys. Res. B* **207**, 45–54 (2003)
- P.J. Blau, B. Dumont, D.N. Braski, Reciprocating friction and wear behavior of a ceramic-matrix graphite composite for possible use in diesel engine valve guides. *Wear* **225–229**, 1338–1349 (1999)
- J.X. Deng, T.K. Cao, X.F. Yang, J.H. Liu, Self-lubrication of sintered ceramic tools with CaF₂ additions in dry cutting. *Int. J. Mach. Tools Manuf.* **46**, 957–963 (2006)
- C.H. Ding, P.L. Li, G. Ran, Y.W. Tian, J.N. Zhou, Tribological property of self-lubricating PM 304 composite. *Wear* **262**, 575–581 (2007)
- A. Gangopadhyay, S. Jahanmir, M.B. Peterson, Self-Lubricating Ceramic Matrix Composites, in *Friction and Wear of Ceramics*, ed. by S. Jahanmir (Marcel Dekker, New York, 1994), pp. 163–197
- M.N. Gardos, H.-S. Hong, W.O. Winter, The effect of anion vacancies on the tribological properties of rutile (TiO_{2-x}), Part II: experimental evidence. *Tribol. Trans.* **22**(2), 209–220 (1990)
- S. Gupta, D. Filimonov, T. Palanisamy, T. El-Raghy, M.W. Barsoum, Ta₂AlC and Cr₂AlC Ag-based composites-new solid lubricant materials for use over a wide temperature range against Ni-based superalloys and alumina. *Wear* **262**, 1479–1489 (2007)
- C.F. Higgs III, E.Y.A. Worniyoh, An in situ mechanism for self-replenishing powder transfer films: experiments and modeling. *Wear* **264**, 131–138 (2008)
- J.H. Ouyang, T. Murakami, S. Sasaki, Y.F. Li, Y.M. Wang, K. Umeda, Y. Zhou, High temperature tribology and solid lubrication of advanced ceramics. *Key Eng. Mater.* **368–372**, 1088–1091 (2008)
- H.R. Pasaribu, J.W. Sloetjes, D.J. Schipper, Friction reduction by adding copper oxide into alumina and zirconia ceramics. *Wear* **255**, 699–707 (2003)
- H.R. Pasaribu, K.M. Reuver, D.J. Schipper, S. Ran, K.W. Wiratha, A.J.A. Winnubst, D.H.A. Blank, Environmental effects on friction and wear of dry sliding zirconia and alumina ceramics doped with copper oxide. *Int. J. Refract. Metals Hard Mater.* **23**, 386–390 (2005)
- A. Skopp, M. Woydt, Ceramic and ceramic composite materials with improved friction and wear properties. *Tribol. Trans.* **38**(2), 233–242 (1995)
- H.E. Sliney, C. Dellacorte, The friction and wear of ceramic/ceramic and ceramic/metal combinations in sliding contact. *Lubr. Eng.* **50**(7), 557–576 (1994)
- L. Tuchinskiy, E. Veksler, R. Loutfy, M. Williams, Tribological characteristics of Si₃N₄-based self-lubricating materials. *Tribol. Trans.* **43**, 603–610 (2000)

M. Woydt, M. Dogli, P. Agatonovic, Concepts and technology development of hinge joints operated up to 1600°C in air. *Tribol. Trans.* **40**(4), 643–646 (1997)

Solid Lubricants, Graphene

ZHIBIN LU¹, LIPING WANG²

¹Lanzhou Institute of Chemical Physics, Chinese Academy of Science, Lanzhou, People's Republic of China

²State Key Laboratory of Solid Lubrication, Lanzhou Institute of Chemical Physics, Chinese Academy of Science, Lanzhou, People's Republic of China

Synonyms

Graphene

Definition

Graphene is a single planar sheet of sp²-bonded carbon atoms. Its extended honeycomb network is the basic building block of three-dimensional graphite, one-dimensional nanotubes, and zero-dimensional fullerenes.

Scientific Fundamentals

Introduction

The rapid progress of solid lubrication technology in the past 4 decades has met the requirements of the aerospace, automotive, and other industries (Miyoshi 2001, Miyoshi et al. 2005). Solid lubricant is used when liquid lubricants do not meet the advanced requirements. In high vacuum (or space), high temperature, low temperature, radiation, clean environment, or corrosive environments, and combinations thereof, solid lubricant may be the only workable system. The materials designed for solid lubrication not only need desirable friction coefficient, but also must maintain good durability in different environments, such as high vacuum, water, air, low temperature, or dust. With the development of micro-/nano-electromechanical systems (MEMS/NEMS), the dimension match need to be considered. Specifically, the separations between MEMS/NEMS can range from 1 μm/100 nm to contact. Since experimentally discovered in 2004 (Novoselov et al. 2004), graphene has exhibited a number of intriguing properties. Theoretical and experimental tests on individual graphene nanosheets exhibit extremely high values of Young's modulus (~1,000 GPa), fracture strength (~125 GPa) (Lee et al. 2008), thermal conductivity (~3,000 W m⁻¹ K⁻¹) (Bolotin et al. 2008), mobility

of charge carriers ($\sim 200,000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) (Matthew et al. 2010), and fascinating transport phenomena such as the quantum Hall effect (Castro Neto et al. 2009). Effective solid lubricants generally have a number of basic properties, such as thermal stability, low friction, low surface adherence, and outstanding elastic and mechanical properties. It is known that carbon-derived materials have outstanding lubrication properties in possession of the above desirable properties. Materials such as graphite have been widely used as lubricants for many years. Graphene is the building block of the common macroscopic solid lubricant graphite, and it is also characterized by the above-mentioned advantages. Based on these results, it can be inferred that graphene would be an effective solid lubricant.

Structure and Electronic Properties of Graphene

Graphene is a flat monolayer of carbon atoms tightly packed into a two-dimensional (2D) honeycomb lattice. Its extended honeycomb network is the basic building block of other important allotropes and can be stacked to form 3D graphite, rolled to form 1D nanotubes, and wrapped to form 0D fullerenes (Fig. 1).

Due to its true 2D structure, graphene has unique properties in many respects that are different from both usual metals and semiconductors (Castro Neto et al. 2009). It has unusual electronic excitations described in terms of Dirac fermions that move in a curved space, is an interesting mix of a semiconductor zero density of states and a metal gaplessness, and has properties of soft matter. The electrons in graphene have very long mean free paths and seem to be almost insensitive to disorder and electron–electron interactions. Graphene has also a robust but flexible structure with unusual phonon modes that do not

exist in ordinary 3D solids. These properties can be easily modified with the application of electric and magnetic fields, addition of layers, control of the geometry, and, chemical doping. Furthermore, graphene can be directly and relatively easily probed by various scanning probe techniques from mesoscopic down to atomic scales, because it is not buried inside a 3D structure.

Elastic and Mechanical Properties of Graphene

The ideal strength is the highest achievable strength of a defect-free crystal at 0 K. Graphene is an ultrastrong material, defined by internal stress levels broadly and persistently rising up to a significant fraction of the ideal strength. Recent developments in density functional theory (DFT) methods applicable to studies of large periodic systems have become essential in addressing problems in material design and processing. The theory allows one to interpret experimental data and to determine the underlying properties of a material that is a well tested ab initio method for accurate mechanical properties. The typical steps are as follows: first, set up a supercell structure, and then apply a series of incremental tensile strains on the supercell and simultaneously relax the other stress components to zero (Poisson contraction under uniaxial tension). Then, the DFT calculated stress and Poisson's ratio varying with finite-deformation can be obtained. A representational study (Liu et al. 2007) showed that the Young's modulus $E = 1,050 \text{ GPa}$ and Poisson's ratio $= 0.186$. In both x and y uniaxial tensions, phonon instabilities occurred near the center of the Brillouin zone, at $\epsilon_{xx} = 0.194$, $\sigma_{xx} = 110 \text{ GPa}$, $\epsilon_{yy} = -0.016$ and $\epsilon_{yy} = 0.266$, $\sigma_{yy} = 121 \text{ GPa}$, and $\epsilon_{xx} = -0.027$. Both soft phonons were longitudinal elastic waves in the pulling direction, suggesting that brittle cleavage fracture would be an inherent behavior of graphene and carbon nanotubes at low temperatures.

Graphene's elastic behavior reflects the intrinsic properties of the interatomic bonding all the way up to the breaking point, because of the low defect density in the films. As suggested by Lee et al. (2008, 2009), the response of the graphene must be considered as non-linear since the stress–strain response must curve over to a maximum point that defines the intrinsic breaking stress. In the simplest model, the resulting isotropic elastic response under uniaxial extension can be expressed as

$$\sigma = E\epsilon + D\epsilon, \quad (1)$$

in which σ is the stress, ϵ the strain, E is the Young's modulus, and D is the third-order elastic modulus. There is a differential operation on both sides of the (1), thus, the maximum of the elastic stress–strain response



Solid Lubricants, Graphene, Fig. 1 The schematics of graphene (middle, upper), fullerene (middle, lower), carbon nanotubes (left), and graphite (right)

defines the intrinsic stress, which for this functional form is $\sigma_{\text{int}} = E^2/4D$ at strain $\varepsilon_{\text{int}} = E/2D$, so it remains only to determine E and D from the experimental results.

Graphene is a truly two-dimensional material, so its strain energy density is normalized by the area of the graphite, rather than by volume normalization. Therefore, its behavior under tensile loading is usually described by a two-dimensional strain σ_{2D} , and elastic constants E_{2D} and D_{2D} , with units of force/length. In order to obtain the corresponding three-dimensional parameters and to compare with bulk graphite and other materials, these quantities can be divided by the interlayer spacing in graphite ($h = 0.335$ nm). Since the sixfold rotation symmetry of the graphene structure, the isotropic mechanical properties are adopted. The force–displacement behavior can be approximated as (Lee et al. 2008, 2009)

$$F = \sigma_0^{2D}(\pi a) \left(\frac{\delta}{a} \right) + E^{2D}(q^3 a) \left(\frac{\delta}{a} \right)^3 \quad (2)$$

where F is the applied force, δ is the deflection at the center point, σ_0^{2D} is the pre-tension in the film, and $q = 1.02$. Therefore, the pre-tension and the elastic modulus can be obtained by fitting the force-displacement curve to (2). For monolayer graphene, the data followed (2) until the large displacement, while the data fell significantly below the data obtained from small-displacement curve fitting for the displacement of more than 50 nm of bilayer graphene. The measured mean values of moduli (E_{2D}) for monolayer, bilayer, and trilayer graphene were 342, 698, and 986 N/m, respectively (Lee et al. 2008, 2009). These correspond to Young's moduli of $E = 1.02$, 1.04, and 0.98 TPa, assuming 0.335 nm as the thickness of one layer. Within experimental error, the Young's moduli of monolayer, bilayer, and trilayer graphene are all identical and equal to the value for bulk graphite.

The fracture strength of graphene was measured by loading the membranes to the breaking point, but the force required to break the membranes depended strongly on the tip radius due to the extreme stress concentration at the tip, which can be described as (Lee et al. 2008, 2009):

$$\sigma_m^{2D} = \left(\frac{FE^{2D}}{4\pi R} \right)^{\frac{1}{2}} \quad (3)$$

where σ_m^{2D} is the maximum stress at the central point of the film, F is the breaking force, and R is the tip radius. Equation (3) yielded an average breaking strength of 55 N/m for monolayer graphene. However, because the model ignored non-linear elasticity, this value overestimated the strength. To extract the true breaking strain, Lee et al. performed a series of finite element simulations, using non-linear stress–strain behavior

given by (1). For monolayer films, this value was $D = 680$ N/m, yielding a two-dimensional ultimate strength of 42 N/m, which corresponded to a strength of 130 GPa in the bulk limit.

To provide a good estimate for bilayer and trilayer membranes, the measured strength of monolayer graphene, together with the tip radius used in each experiment, was used to scale the multilayer results. The estimated strength, which represented an upper bound of the true strength, decreased with the thickness, from 126 GPa for bilayers to 101 GPa for trilayer graphene (Lee et al. 2008, 2009).

Frictional Properties of Graphene

Generally, the microtribological behavior was evaluated using an AFM. The friction forces are proportional to the applied loads obeying Amonton's law

$$F_L \cdot \dot{y} = \mu F_N + F_0 \quad (4)$$

where μ is the friction coefficient, F_L is the friction force, F_N is the normal load, and F_0 is the friction force at an external load of zero. Actual friction forces are difficult to obtain, and F_L is expressed as voltage signals; correspondingly, the real friction coefficients ($\mu = \partial F_L / \partial F_N$) could not be feasibly calculated. However, a relative friction coefficient (RFC), which is the slope of the force curve, can be obtained. Therefore, for a given AFM probe, the RFCs for various samples can be compared with each other.

The results of micro-scale frictional experiments performed in an ambient environment (Lee et al. 2009) showed that, independent of scan speed (from 1 to 10 mm/s) and the applied load (from 0.1 to 2 nN), the friction force decreases monotonically with sample thickness, and converges to that of bulk graphite as the number of layers increase above 5.

Atomic-scale frictional experiments tested in a dry nitrogen purged chamber by Lee et al. showed that the energy dissipation decreased monotonically as the number of graphene layers increases, and approaches that of bulk graphite, consistent with the micrometer-scale measurements in ambient conditions. In addition, the adhesion force between the tip and graphene flakes with different numbers of layers showed that there was virtually no difference between different thicknesses. Furthermore, measured friction on suspended graphene membranes showed that there is no difference between suspended and supported graphene.

After ruling out substrate effects and influence of adhesion, Lee et al. concluded that the layer-dependence of friction appears to be governed by dissipation mechanisms taking place at the tip-graphene interface.

Understanding the fundamental origin of friction is one of the pressing challenges with the miniaturization of moving components in many commercial products, including computer disk heads and the microelectromechanical systems that trigger the airbags in cars. At the core of the problem is an identification of the mechanisms that convert kinetic energy of sliding contacts into heat. By reducing the thickness of the solid lubricant graphite to the most extreme limits, single and double atomic layers of graphene grown epitaxially on SiC, Filleter et al. demonstrated that there was a significant difference in friction between single and bilayer graphene films. And this difference was related to dramatic reduction in electron-phonon coupling in the bilayer as revealed by angle-resolved photoemission spectroscopy. Filleter et al. explored friction and dissipation in epitaxial graphene films, revealing that bilayer graphene as a lubricant outperforms even graphite due to reduced adhesion (Filleter et al. 2009).

Tribological Properties of Reduced Graphene Oxide Sheets

The characteristic properties of graphene include outstanding thermal conductivity, novel electronic properties, and excellent mechanical behavior and make it the ideal candidate for the basic building block of MEMS/NEMS. However, the difficulty of obtaining a stable graphene structure on device surfaces has hindered its potential in MEMS/NEMS because of the lack of function groups in graphene. Because of abundance of oxygenous groups in graphene oxide (GO) molecules, it is advisable to assemble the reduced GO (RGO) microstructure on a substrate by performing thermal treatment.

Following the above, Ou et al. (2010) have assembled the stable RGO sheets onto the surface of a silicon wafer by a multi-step method. First, a 3-aminopropyl triethoxysilane self-assembled monolayer (coded APTES-SAM) was prepared on the silicon wafer via a self-assembly process. Then GO was grafted onto a silicon substrate covered with APTES-SAM through chemical reactions between the oxygenous and amine groups. Finally, the RGO was obtained by annealing the formed GO, in which the obtained sample was coded APTES-RGO. WCA measurement is a simple, useful, and sensitive tool for gaining insight into surface chemical components. Ou et al. have shown that the hydroxylated silicon wafer and APTES-GO were hydrophilic with WCA of ~ 0 and 40.2, respectively. Correspondingly, high adhesive forces of $\sim 190/\sim 140$ nN were generated between the AFM tip and the samples of hydroxylated silicon substrate/APTES-GO. Once APTES-GO was reduced, WCA

increases to 85.5, and adhesive force decreased to ~ 100 nN, exhibiting good adhesion resistance.

The microtribological properties evaluated with an AFM by Ou et al. show that the hydroxylated silicon wafer possessed the highest friction and RFC. Both APTES-GO and APTES-RGO reduced friction. APTES-RGO exhibited the best lubricity, which could be assigned to the smallest adhesive force compared with the hydroxylated silicon wafer and APTES-GO as well as the lowest surface energy (reflected by the highest WCA of 85.5).

The macrotribological behavior of APTES-SAM, graphite, APTES-GO, APTES-RGO, and RGO on bare Si substrate at an applied load of 0.1 N, and APTES-RGO at an applied load of 0.2 N, have been tested on a ball-on-plate macrotribometer by Ou et al. Test results show that APTES-SAM displayed poor tribological properties characterized by high friction and very short antiwear life under the testing conditions of 0.1 N and 1 Hz. Once GO was assembled, the sample exhibited improved tribological properties characterized by a reduced friction coefficient (~ 0.25) and a lengthened antiwear life ($\sim 1,600$ s). Upon thermal reduction, antiwear life increased further to $>10,800$ s and the reduced friction coefficient remained unchanged.

The friction coefficient of the APTES-RGO was a little lower than that of graphite plate (about 0.28) and lower than the friction coefficient of carbon nanotube (CNT) (about 0.3–0.5) tested by a similar ball-on-disk tribometer.

Key Application

Tribological Application of Composite Coating and Composites Reinforced with Graphene

Because of the effects of their improved mechanical properties and unique structure, carbon-derived materials such as fullerene and carbon nanotubes can be used to fabricate composites and coatings with excellent tribological performance attributed to their excellent mechanical properties and the unique topological structure of carbon-derived materials.

Graphene is considered an excellent candidate for developing functional and structure-reinforced composites because of its remarkable mechanical properties, such as exceptionally high elastic modulus, large elastic strain, and fracture strain sustaining capability. Incorporation of graphene into polymer matrices remarkably improves the mechanical properties of the materials (Wang et al. 2009, Zhao et al. 2010). It is expected that graphene-reinforced

composites and coatings will have extensive tribological application. The introduction of graphene can improve the mechanical properties of composites and coatings. In addition, the formation of carbon film can reduce the friction and wear rate.

Graphene Platelets as a Lubricant Additive

Lin et al. (2011) demonstrated that the graphene platelets modified by stearic and oleic acids could be added as lubricant additive to stably disperse in oil. The wear resistance and load-carrying capacity of the lubricating oil were greatly improved with the addition of the modified graphene platelets at an optimal content of 0.075 wt.%. During the friction process, the wear rate and the friction coefficients of both the base oil and the oil with the modified graphene platelets increased remarkably. However, the wear rates and the friction coefficient of the oil with modified graphene platelets were relatively steady throughout the test. Moreover, the friction coefficient of the oil with modified graphene platelets was much lower than that of the base oil and oil with modified graphene platelets. And the modified graphene platelets have better tribological properties than the modified graphene platelets when used as additives.

Graphene/RGO as Lubricant on MEMS/NEMS

A graphene/RGO sheet, which has a thickness of several nanometers, matched with the separations between MEMS/NEMS, will be an ideal choice as lubricant film on a nanoscale due to not only the excellent electronic and mechanical properties but also because of potential friction reduction and antistiction performance.

Cross-References

- Graphite Solid Lubrication Materials
- Solid Lubricants

References

- K.I. Bolotin, K.J. Sikes, Z. Jiang, M. Klima, G. Fudenberg, J. Hone, P. Kim, Ultrahigh electron mobility in suspended graphene. *Solid State Commun.* **146**, 351–355 (2008)
- A.H. Castro Neto, F. Guinea, N.M.R. Peres, K.S. Novoselov, A.K. Geim, The electronic properties of graphene. *Rev. Mod. Phys.* **81**, 109–162 (2009)
- T. Filletier, J.L. McChesney, A. Bostwick, E. Rotenberg, K.V. Emtsev, Th Seyller, K. Horn, R. Bennewitz, Friction and dissipation in epitaxial graphene films. *Phys. Rev. Lett.* **102**, 086102 (2009)
- C. Lee, X. Wei, J.W. Kysar, J. Hone, Measurement of the elastic properties and intrinsic strength of monolayer graphene. *Science* **321**, 385–388 (2008)
- C. Lee, X. Wei, Q. Li, R. Carpick, J.W. Kysar, J. Hone, Elastic and frictional properties of graphene. *Phys. Status Solidi B* **246**, 2562–2567 (2009)

- J. Lin, L. Wang, G. Chen, Modification of graphene platelets and their tribological properties as a lubricant additive. *Tribol. Lett.* **41**, 209–215 (2011)
- F. Liu, P. Ming, J. Li, *Ab initio* calculation of ideal strength and phonon instability of graphene under tension. *Phys. Rev. B* **76**, 064120 (2007)
- J.A. Matthew, C.T. Vincent, B.K. Richard, Honeycomb carbon: a review of graphene. *Chem. Rev.* **110**, 132–145 (2010)
- K. Miyoshi, *Solid Lubrication Fundamentals and Applications* (Marcel Dekker, New York, 2001)
- K. Miyoshi, K.W. Street Jr., R.L. Vander Wal, A. Rodney, A. Sayir, Solid lubrication by multiwalled carbon nanotubes in air and in vacuum. *Tribol. Lett.* **19**, 191–201 (2005)
- K.S. Novoselov, A.K. Geim, S.V. Morozov, D. Jiang, Y. Zhang, S.V. Dubonos, I.V. Grigorieva, A.A. Firsov, Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004)
- J. Ou, J. Wang, S. Liu, B. Mu, J. Ren, H. Wang, S. Yang, Tribology study of reduced graphene oxide sheets on silicon substrate synthesized via covalent assembly. *Langmuir* **26**, 15830–15836 (2010)
- D.W. Wang, F. Li, J.P. Zhao, W.C. Ren, Z.G. Chen, J. Tan, Z.S. Wu, I. Gentle, G.Q. Lu, H.M. Cheng, Fabrication of graphene/polyaniline composite paper via in situ anodic electropolymerization for high-performance flexible electrode. *ACS Nano* **3**, 1745–1752 (2009)
- X. Zhao, Q. Zhang, D. Chen, P. Lu, Enhanced mechanical properties of graphene-based poly (vinyl alcohol) composites. *Macromolecules* **43**, 2357–2363 (2010)

Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides

TOMAS POLCAR

Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic

Definition

Transition metal dichalcogenides (TMD) are among the best alternatives as solid lubricants for tribological applications, particularly in dry and vacuum environments. Their excellent lubricating properties are based on the extreme crystal anisotropy; under ideal conditions (very dense material deposited and tested in ultra-high vacuum) they can be considered as “frictionless,” i.e., with friction coefficient as low as 0.001. Pure TMD are sensitive to the presence of water vapor, exhibiting high friction coefficient. For tribological purposes, TMD are prepared in the form of powder, burnished films, thin films, inorganic fullerene-like particles, or nanotubes.

History and Fundamental Characterization of Transition Metal Dichalcogenides

Transition metal dichalcogenides (TMD) are, in many ways, a gift from Nature to mechanical engineers looking for friction reduction. The TMD family consists of molybdenum, tungsten and niobium disulfides, and diselenides (Bergmann et al. 1981). From a tribological point of view, tellurides, which technically belong to chalcogenides as well, are not attractive. Only molybdenum disulfide (molybdenite) and tungsten disulfide (tungstenite) occur naturally; the former in abundance, while the later only as a very rare mineral. They are typically contaminated by carbon and oxygen; moreover, tungstenite often contains molybdenum as well.

It is probable that the first report describing molybdenum disulfide (“molybdena”) was given by Pliny the Elder; however, the first definition distinguishing molybdenite from graphite was published by J.A. Cramer in 1764. The first tribological applications of MoS₂ were exclusively connected with the oil industry, with dispersed MoS₂ powder in oils and greases (1938). In 1941, MoS₂ was used with great success as a solid lubricant in high vacuum conditions. In the late 1970s, the sputtering technique was used to deposit thin MoS₂ film later applied to satellites and spacecrafts. Today, TMD is often used either as an oil additive or as a coating.

TMD exist in two crystal forms, hexagonal and rhombohedral. Only the hexagonal structure will be discussed, since it is the most common and important for low-friction applications. The hexagonal crystal structure with sixfold symmetry exhibits a laminar structure. Each sulfur atom is equidistant from three metal atoms, and each metal atom is equidistant from six dichalcogenide atoms. Large spacing between C-M-C (C – chalcogenide, M – transition metal) layers and weak van der Waals forces facilitate easy inter- or intra-crystalline slip (Lansdown 1999). It has been shown that six non-bonding electrons completely fill a band that confines the electron within the crystal structure. Consequently, a net positive charge is created on the surface of C-M-C layers, causing electrostatic repulsing and resulting in easy shear. On the other hand, strong intra-planar covalent bonding helps resist asperity penetration, even under extremely high contact pressures. Thanks to the layered structure, the properties of TMD are highly anisotropic. The hardness of the natural TMD minerals is 1–1.5 on the Mohs scale.

Sulfur-based TMD exhibit a very good thermal stability in an inert gas with sulfur release at temperatures close to 1,000°C; it is expected the selenium is released from selenium-based TMD at lower temperatures. The oxidation resistance of TMD is generally low. The

oxidation of molybdenum and tungsten disulfides starts at 400°C and 500°C, respectively, whereas the oxidation resistance of diselenides is significantly lower. The presence of water is detrimental for any TMD due to induced oxidation, even at room temperature. The oxidation is a complex process depending on environmental and contact conditions; the typical ultimate products of the oxidation are molybdenum or tungsten trioxide.

Scientific Fundamentals

Lubrication Mechanism

The shear between adjacent lamellae of TMD is easy due to low inter-lamellar forces represented only by weak van der Waals forces. High inter-lamellar spacing is attributed to the lack of electronic interactions. The inter-lamellar bonding within a crystal is at a minimum and the presence of contaminants thus hinders lubrication properties due to increasing inter-lamellar interaction. Thus, TMD fundamentally differ from other well-known layered materials such as mica or graphite with strong ionic bonding. To decrease friction of graphite, the bond energies must be reduced by the presence of contaminants, typically water. As a result, the friction of graphite in dry ambient or vacuum is high, whereas TMD exhibits higher friction in humid atmosphere. Despite contradictory reports, it seems that there is no fundamental difference between interfacial and inter-crystalline friction of TMD.

The friction of TMD decreases with applied load and deviates from Amonton's law. Shear stress of solids at high pressures could be approximated as

$$S = S_0 + \alpha P \quad (1)$$

where S is the shear stress at pressure P , S_0 is the shear stress at zero applied pressure, and α is a constant. Dividing throughout by P , the coefficient of friction μ is

$$\mu = \frac{S_0}{P} + \alpha \quad (2)$$

In well-defined contact geometries the pressure P could be calculated from the Hertzian contact model provided the substrates are loaded below the elastic limit. The friction for ball-on-disc contact is then

$$\mu = S_0 \pi \left(\frac{3R}{4E} \right)^{\frac{2}{3}} W^{-\frac{1}{3}} + \alpha \quad (3)$$

By changing load, (3) can be fitted, giving shear stress at zero applied pressure and α . Considering a very low value of α (typically around 0.001 ± 0.001 for MoS₂), plots of $\log(\mu)$ against $\log(W)$ can be easily examined.

There are a number of studies showing linear dependence of $\log(\mu)$ vs $\log(W)$ with slope close to -0.33 , thus validating (3). Moreover, the validity of (3) strongly supports the view that the low friction of TMD is due to shear between lamellae and not by cleavage, as suggested by older theories.

Generally there are three conditions to be satisfied to achieve low friction TMD properties: (1) transfer of TMD on the counterpart, (2) absence of contaminants, and (3) orientation of basal planes parallel to surface.

Transfer of the TMD occurs during initial contact with the counterpart. Thanks to high adhesion of TMD on most of surfaces and low weak bonding between lamellae, the TMD layer is formed very quickly. The coverage of the counterpart by the transfer film is enhanced by the high proportion of edge-sites of lamellae facilitating further building of the transfer layer. Transfer layer formation is also influenced by the initial crystal orientation of TMD; well-oriented TMD forms a tribolayer almost immediately. In real tribological contact, the transfer could be direct or through attachment of worn TMD particles. The thickness of the transferred layer varies from nanometers to tens of microns.

As already noted, the basal plane orientation of the TMD grains is a key factor for low friction. Fortunately, the friction-induced change of orientation of basal planes is a rapid process and running-in is very short; high contact pressure facilitates re-orientation, which occurs simultaneously with the transfer of TMD material on the counterbody. Several models of re-orientation have been discussed, such as fracture or removal of wrongly oriented crystals and their reattachment with optimum basal plane orientation, plastic bending, or chemical processes induced by sliding. It seems that models stick well to individual material and sliding conditions; however, they cannot be generalized.

The sensitivity of TMD to the presence of contaminants significantly limits their use as low friction material in many, particularly terrestrial, applications. Typical contaminants are oxygen-forming molybdenum oxides or substituting sulfur atoms in the lamellar structure, carbon, or water vapor condensing in the defects of TMD crystal structure. Consequently, the lowest friction coefficient could be achieved only in ultra-high vacuum (UHV) conditions; TMD must be absolutely pure and stoichiometric. In such conditions, Martin et al. reported superlubricity due to frictional anisotropy of the shear-oriented low-energy basal planes when sliding MoS_2 in UHV against steel counterpart (Martin et al. 1993). In some cases, the frictional force was below the resolution of the equipment.

At terrestrial atmosphere with relative humidity close to 50%, molybdenum and tungsten disulfides exhibit friction in the range 0.1–0.2; the friction of selenides is slightly lower. In general, the friction increases with air humidity. Compared with water vapor, the effect of air oxygen is less dramatic, leading only to limited increase of the friction. The decrease of friction with increase of temperature is due to drying of air humidity; the friction is then very low, with values close to that of sliding in dry air.

Preparation of TMD for Tribology Applications

Today, TMD for tribological purposes are mostly prepared as a thick films (e.g., burnished), films prepared by electrochemical processes, powders as oil additives, or a thin films deposited mainly by physical vapor deposition (PVD). TMD could be formed as well during the sliding process on the surfaces in the contact by tribochemical reactions (Grossiord et al. 1998). Recently, attention has been paid to fullerene-like TMD (Rapoport et al. 1997) and TMD nanotubes (Remskar et al. 2001), both as additives and films.

The simplest way to prepare burnished TMD film is to apply a sliding pressure to TMD powder against surface to be coated. With rougher substrates, TMD tends to fill in low spots on the surface and thus produce a smoother surface. When full coverage is achieved, the TMD film grows, keeping an identical texture (Lansdown 1999). This is beneficial for many applications, since the final roughness of burnished films is (to some extent) independent of the initial one. Burnished film consists of randomly oriented TMD crystals except for the utmost surface, where the basal planes are parallel to the surface.

PVD methods are the most versatile for preparing dense and well-adhered TMD coatings with different microstructures; magnetron sputtering is the deposition process that is most often used on an industrial scale. Properties of sputtered films can be tailored by tuning of deposition parameters, such as substrate bias, working pressure, target-to-substrate distance, substrate temperature, and so on. Typical sputtered TMD films exhibit columnar morphology and a mixture of phase orientations; the films are often porous. With regard to chemical composition, the films are metal-rich due to re-sputtering of chalcogenide atoms from the target surface and chemical reaction of chalcogenides with residual atmosphere. However, it has been shown that TMD retains low-friction behavior when with metal-to-chalcogenide ratio is higher than 1.2. There has been great effort made to deposit TMD with (002) orientation; however, when the thickness of such layer reached

tens of nanometers, (102) orientation became dominant. Recently, pulsed magnetron sputtering discharges were have been used to control MoS₂ basal plane orientation during film growth (Muratore and Voevodin 2009).

Nevertheless, pure sputtered TMD films have following drawbacks: (1) low hardness and thus low resistance to abrasion, (2) low load-bearing capacity, and (3) high sensitivity to water vapor. One of the most promising and frequently used solutions is co-sputtering of TMD with additional element(s) or compounds, which will be discussed next.

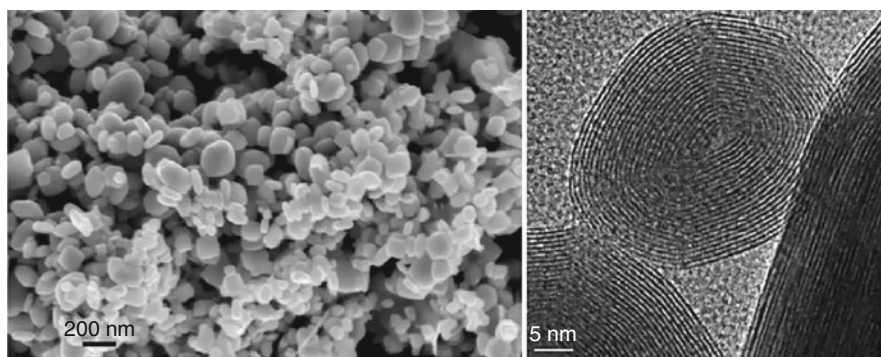
TMD nanotubes (NT) can be prepared by many methods, such as thermal decomposition, chemical transport, gas phase, or solid-gas reactions; either single-wall or multi-wall NT can be prepared. Compared with carbon NT, it is much more difficult to keep homogeneous distribution of diameter and length during the preparation process. They were tested as oil additives with concentration up to 1 wt%. Despite some sheets demonstrating exfoliation, intact NT were found in the tribofilm. It seems that TMD NT resist the sliding process well, experiencing only few structural modifications (Martin and Ohmae 2008). Sliding in oil with dispersed NT did not exhibit any significant advantage compared with standard lubrication additives.

Inorganic fullerene-like nanoparticles, IF-TMD, can be described as hollow spheres fitted together, forming an onion-like structure (Fig. 1). Thanks to their shape, IF-TMD exhibit high thermal stability and oxidation resistance. Typically, they are synthesized by gas phase reaction (MoS₂-based) or solid-gas reaction (WS₂-based). IF-TMD were first tested as an oil additive. Two possible low-friction mechanisms have been proposed – bearing effect, with fullerenes as nanobearings, and exfoliation (i.e., the sliding mechanism similar to lamellar TMD). There has been an attempt to demonstrate that

the bearing effect exists and that friction and wear are dependent on the IF-TMD dimension; however, the agglomeration of nanoparticles, unexpected transferred layer composition, and unique elastic properties of IF-TMD make it difficult to draw any definite conclusions (Martin and Ohmae 2008). On the other hand, there is no doubt that low friction of IF-TMD is related to formation of a layered structure in the contact area. It has been shown that the improved solid-state lubrication properties of IF-TMD are extrinsic, i.e., they do not depend on the elastic properties of particles. The main role of the nanoparticles is to provide TMD material for contact layer through deformation and delamination; the nanoparticles may fill the gaps on the surface and act as solid-lubricant reservoirs (Brown et al. 2007). IF-TMD are intensively studied, mainly as oil additives; however, despite several successful laboratory tests, their industrial applicability is still very limited.

TMD-Based Thin Films Alloyed with Elements or Compounds

The first attempts to modify TMD films by alloying were aimed to increase density and consequently reduce porosity, improve adhesion, and significantly increase of the hardness. In general, the aims have been achieved. Among many metals for alloying (Ti, Al, Au, Pb, Ni, Cr), titanium was probably the most successful from the commercial point of view (MOST[®] by Teer, Ltd.; MoS₂ alloyed with Ti). The improvement of the mechanical properties was evident; however, Ti addition could explain the increase in the wear resistance and the load-bearing capacity but not the fact that low friction was observed. It has been speculated that Ti reduced the oxidation of TMD films by (1) a getter effect during the deposition and (2) preferential oxidation of the film compared to the TMD (Teer 2001). However, new results do not support



Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides, Fig. 1 SEM micrograph of MoS₂ fullerene-like particles and HR-TEM micrograph showing structure of individual particle (Reprint from Rosentsveig et al. 2009)

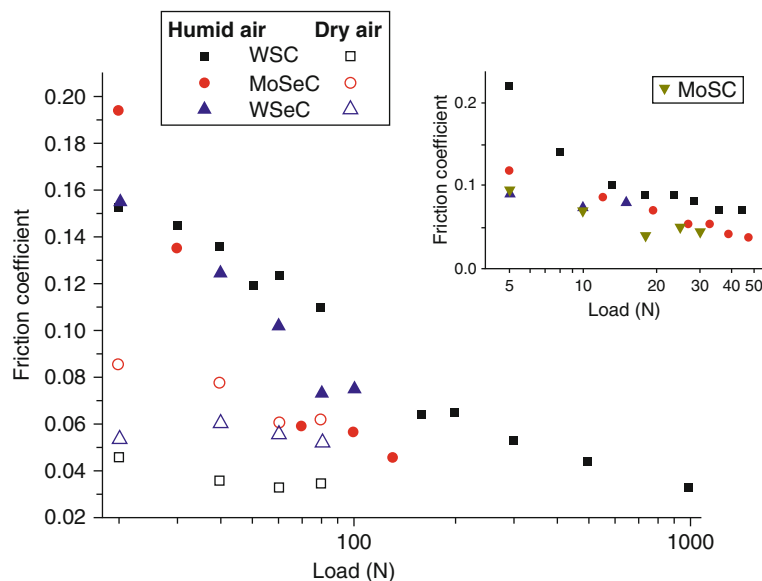
such a hypothesis (Scharf et al. 2009). Moreover, the alloying with “inert” metals, such as Au, resulted in similar sliding properties. Alloying with compounds such as ZnO, Sb₂O₃, or PbO was analyzed in the 1990s; however, the benefit was negligible compared with that of metals. All films referred to above show two general features: prevalence of the TMD phase (maximum content of alloying is about 20 at.%) and limited interaction between the TMD phase and the alloying element (contamination of oxygen and carbon from the residual atmosphere during the sputtering process is not considered here as alloying).

In the late 1990s, a new concept of coatings based on the alloying of transition metal dichalcogenides (TMDs) with carbon started to attract the attention of several scientific groups. The original idea was to combine the excellent frictional behavior of TMDs in vacuum and dry air with the tribological properties of DLC coatings. Moreover, an increase in the coating compactness in relation to TMD and an improvement in the mechanical properties, particularly the hardness, was expected. Zabinski et al. (Voevodin and Zabinski 2000) prepared W-S-C coatings either by magnetron-assisted pulsed laser deposition (MSPLD – target WS₂) or by laser ablation of a composite target made of graphite and WS₂ sectors. The coatings showed nanocomposite structure combining tungsten carbide and tungsten disulfide

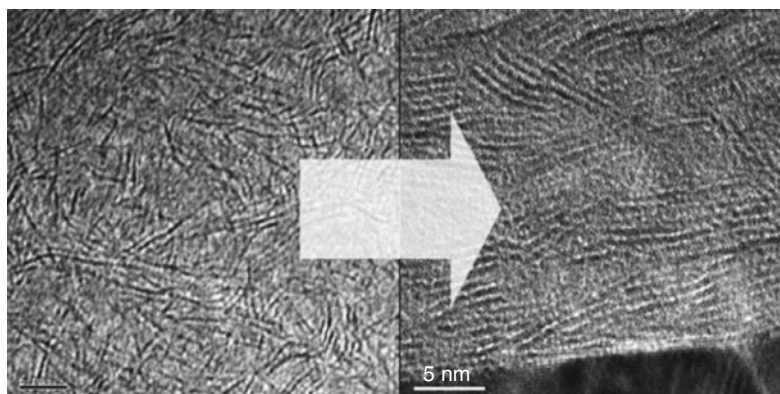
nanocrystals in the carbon matrix. The friction coefficient in dry air was lower than the one measured in humid air (0.02 and 0.15, respectively); however, environmental cycling (i.e., humid air/dry nitrogen) showed “chameleon behavior” with increasing and decreasing friction as a function of atmosphere. Such behavior was attributed to change of sliding mechanism, which was driven by WS₂ in dry air and by carbon in humid air. W-S-C coatings prepared by magnetron sputtering have been intensively studied by Cavaleiro et al. The analysis of the microstructure of the reactive deposited W-S-C coatings (Nossa and Cavaleiro 2004) shows that, for high carbon content, they are formed of a mixture of nanocrystals of WS₂ phase side-by-side with nanocrystals of W-C phases and small C-based amorphous zones, or just by WS₂ nanograins dispersed in an amorphous carbon matrix (low carbon content).

Finally, a series of TMD co-deposited with carbon exhibited unique microstructure with randomly oriented separated TMD platelets in the carbon matrix. The coatings showed extremely high load-bearing capacity, low sensitivity to air humidity, and high wear resistance (Polcar et al. 2009), see Fig. 2.

Excellent tribological properties of TMD-C film are attributed to adaptation of coating microstructure to sliding process (Fig. 3). Optimally oriented TMD platelets



Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides, Fig. 2 Friction coefficient of TMD-C coatings as a function of the applied load in dry and humid conditions, SRV test, 20,000 laps, 100Cr6 steel ball with a diameter of 10 mm. Insets shows the friction coefficient of TMD-C coatings measured on pin-on-disc in humid air, 5,000 laps, 100Cr6 steel ball with a diameter of 6 mm



Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides, Fig. 3 Randomly oriented MoSe₂ platelets (left) in as-deposited MoSeC coatings are reoriented after sliding process (right)

form a very thin (up to 5 nm) tribolayer on the surface. The reorientation of the TMD platelets inside the carbon matrix is then done progressively from some tens of nanometers until the track surface (Polcar et al. 2008).

Future Perspectives

TMD are very promising solid lubricants for MEMS (Scharf et al. 2006), where they simultaneously act as lubricant and surface separation (adhesion forces are extremely high compared with dynamic ones). Two-dimensional TMD nanosheets produced by liquid exfoliation (Coleman et al. 2011) could be used for MEMS as well; moreover, they could be used as oil additives. Advanced deposition methods using high power impulse magnetron sputtering (HIPIMS) should lead to unique TMD coating microstructures with improved stoichiometry, density, and adhesion.

Key Applications

Powder MoS₂ is used for cold forming and dry lubrication. TMD particles are used as an additive to oils and greases. TMD thin films are widely used as solid lubricant in vacuum and space mechanisms, such as bearing races, telescopic booms, and robotic joints. Even at low orbit with the presence of highly reactive atomic oxygen, MoS₂ films showed excellent tribological performance. MoS₂ and WS₂ films are often used as a top layer on hard protective coatings, particularly in for tools used in dry cutting conditions. Composite coatings based on TMD are applied in stamping tools, dies, or moulds, where they decrease friction and thus energy consumption, and act as an anti-fouling surface treatment. The automotive industry has become a key market for TMD-based films (mainly piston rings).

Cross-References

- [Amonton's Laws of Friction](#)
- [Chameleon or Smart Solid Lubricating Coatings](#)
- [MoS_x Coatings by Closed-Field Magnetron Sputtering](#)
- [Solid Lubricants for Space Mechanisms](#)

References

- E. Bergmann, G. Melet, C. Müller, A. Simon-Vermot, Friction properties of sputtered dichalcogenide layers. *Tribol. Int.* **14**, 329 (1981)
- S. Brown, J.L. Musfeldt, I. Mihut, J.B. Betts, A. Migliori, A. Zak, R. Tenne, Bulk vs Nanoscale WS₂: finite size effects and solid-state lubrication. *Nano Lett.* **7**(8), 2365 (2007)
- J.N. Coleman, M. Lotya, A. O'Neill, S.D. Bergin, P.J. King, U. Khan, K. Young, A. Gaucher, S. De, R.J. Smith, I.V. Shvets, S.K. Arora, G. Stanton, H.-Y. Kim, K. Lee, G.T. Kim, G.S. Duesberg, T. Hallam, J.J. Boland, J.J. Wang, J.F. Donegan, J.C. Grunlan, G. Moriarty, A. Shmeliov, R.J. Nicholls, J.M. Perkins, E.M. Grieveson, K. Theuvsen, D.W. McComb, P.D. Nellist, V. Nicolosi, Two-dimensional nanosheets produced by liquid exfoliation of layered materials. *Science* **331**, 568 (2011)
- C. Grossiord, K. Varlot, J.-M. Martin, Th LeMogne, C. Esnouf, K. Inoue, MoS₂ single sheet lubrication by molybdenum dithiocarbamate. *Trib. Int.* **31**, 737 (1998)
- A.R. Lansdown, *Molybdenum Disulphide Lubrication* (Elsevier, Amsterdam, 1999)
- J.M. Martin, N. Ohmae (eds.), *Nanolubricants* (Wiley, Chichester, 2008)
- J.M. Martin, C. Donnet, Th LeMonge, Th Epicier, Superlubricity of molybdenum disulphide. *Phys. Rev.* **48**, 10583 (1993)
- C. Muratore, A.A. Voevodin, Control of molybdenum disulfide basal plane orientation during coating growth in pulsed magnetron sputtering discharges. *Thin Solid Films* **517**, 5605 (2009)
- A. Nossa, A. Cavaleiro, Chemical and physical characterization of C(N)-doped W-S sputtered films. *J. Mater. Res.* **19**, 2356 (2004)
- T. Polcar, M. Evaristo, R. Colaço, C. Silviu Sandu, A. Cavaleiro, Nanoscale triboactivity: the response of Mo-Se-C coatings to sliding. *Acta Mater.* **56**, 5101 (2008)
- T. Polcar, M. Evaristo, A. Cavaleiro, Comparative study of the tribological behavior of self-lubricating W-S-C and Mo-Se-C sputtered coatings. *Wear* **266**, 388 (2009)

- L. Rapoport, Y. Bilik, Y. Feldman, M. Homyonfer, S.R. Cohen, R. Tenne, Hollow nanoparticles of WS₂ as potential solid-state lubricants. *Nature* **387**, 791 (1997)
- M. Remskar, A. Mrzel, Z. Skraba, A. Jesih, M. Ceh, J. Demsar, P. Stadelmann, F. Levy, D. Mihailovic, Self-assembly of subnanometer-diameter single-wall MoS₂ nanotubes. *Science* **292**, 479 (2001)
- R. Rosentsveig, A. Gorodnev, N. Feuerstein, H. Friedman, A. Zak, N. Fleischer, J. Tannous, F. Dassenoy, R. Tenne, Fullerene-like MoS₂ nanoparticles and their tribological behavior. *Tribol. Lett.* **36**, 175 (2009)
- T.W. Scharf, S.V. Prasad, M.T. Dugger, P.G. Kotula, R.S. Goeke, R.K. Grubbs, Growth, structure, and tribological behavior of atomic layer-deposited tungsten disulphide solid lubricant coatings with applications to MEMS. *Acta Mater.* **54**, 4731 (2006)
- T.W. Scharf, A. Rajendran, R. Banerjee, F. Sequeda, Growth, structure and friction behavior of titanium doped tungsten disulphide (Ti-WS₂) nanocomposite thin films. *Thin Solid Films* **517**, 5666 (2009)
- D.G. Teer, New solid lubricant coatings. *Wear* **251**, 1068 (2001)
- A.A. Voevodin, J.S. Zabinski, Supertough wear-resistant coatings with 'chameleon' surface adaptation. *Thin Solid Films* **370**, 223 (2000)

Solid Lubricants, Polymer-Based Self-Lubricating Materials

V. N. ADERIKHA¹, A. P. KRASNOV²

¹Technology of Polymer Composite Materials and Parts, V.A. Belyi Metal Polymer Research Institute, National Academy of Sciences of Belarus, Gomel, Belarus

²N.A. Nesmeyanov Institute of Organoelement Compounds, Russian Academy of Sciences, Moscow, Russia

Synonyms

[Antifricition materials and composites](#); [Friction-reducing additives](#); [Internally lubricated compounds](#); [Low friction materials](#)

Definitions

Lubricant is a material introduced onto the friction surface to reduce the friction force and (or) the wear rate. Lubricity is the ability of a material to reduce wear and the friction force irrespective of its viscosity. (GOST 27674–88. Friction, wear and lubrication. Terms and definitions.) Lubricity is the property of forming a lubricating film between moving surfaces, particularly when such surfaces are subject to heavy loads and rapid movements (Campbell et al. 1966). Self-lubricity of polymers and polymer-based materials is a property that ensures operation of a friction unit in absence of external lubrication. Polymer-based

self-lubricating materials (SLM) comprise pure polymers with high intrinsic lubricity, i.e., polytetrafluoroethylene (PTFE), linear polyethylenes (PE), and composite materials based on other polymer binders in which lubricity is supplemented by introduction of either lubricants or nano-sized additives.

Scientific Fundamentals

A great number of substances have been selected, based on practice, as lubricants, both organic, such as soaps, fats, waxes, certain polymers (PTFE, HDPE), and inorganic ones – natural and synthetic crystalline solids, mostly of lamellar structure (graphite and various graphite-based products like fluorinated graphite, thermally expanded graphite, graphene, metal dihalcogenides, molybdenum disulfide (MoS₂), hexagonal boron nitride, talc, carbon fiber, etc.), soft metals and their alloys (indium, tin, lead, silver, gold, bronze, brass, babbitt), metal salts (sulfides, selenides, chlorides, iodides, oxides, hydroxides), and glasses (boron oxide, silicates, phosphates). Recently, the list was supplemented with nano-sized compounds and structures of carbon (fullerenes, tubes, onions, diamonds, fibers) and of other substances (nitride and oxide ceramics, nano-MoS₂, nano-BN) able to efficiently reduce friction of greases, solid lubricating coatings, and polymer-based composites. Solid lubricating coatings and non-polymeric lubricants are not reviewed here, being a subject of separate entry (see ► [Solid lubricants, Layered-Hexagonal Transition Metal Dichalcogenides](#)).

The Mechanism of Self-lubrication of Polymers

A number of neat polymers display low friction on steel and other hard surfaces with friction coefficients comparable to those of classic solid lubricants (graphite, molybdenum disulfide), and some polymers even surpass them in lubricity. These include linear polyolefins, PTFE, and HDPE and its supermolecular weight analogue, the ultrahigh molecular weight polyethylene UHMWPE, which show friction coefficients of 0.03–0.1 (PTFE) and 0.08–0.16 (UHMWPE) coupled with good wear resistance under certain test conditions (low speed, high load, ambient temperature, counter-surface roughness $R_a \sim 0.1 \mu\text{m}$).

Polyoxymethylene (POM), polyamides (PA), polymethylmethacrylate (PMMA), polyetheretherketone (PEEK), and some other semicrystalline and amorphous polymers, distinguished for self-lubricity, generally show higher friction and wear rates in a narrower range of test conditions, beyond which both friction and wear rate increase significantly. The property of many polymers to

enable operation of a friction unit in absence of external lubrication testifies to their self-lubricity. Self-lubricity is a result of mechanical (elastic and plastic deformation, microcutting, shear), physical (orientation, melting), and tribochemical (degradation, cross-linking) processes in the friction zone, producing a transfer film on the counter-surface, modified friction layer on a polymer sample, and wear debris in the clearance, composing together the “third body,” which determines the level of friction losses in the particular friction couple. The transfer films at friction of neat polymers result from modification of the counterbody surface with tribo-degradation products similar in structure and composition to the original polymer.

The typical self-lubricating polymers, the linear polyolefins, possess valuable tribological characteristics owing to their regular chemical structure, i.e., the absence of bulky side groups, substitutes, and branches, which endows them with the so-called “smooth molecular profile” (Pooley and Tabor 1972). An important feature is the absence of reactive or polar groups in the main chain and high bonding energy of C-F and C-H bonds, providing for high chemical resistance, weak intermolecular interaction, and low density of the cohesive energy, facilitating shear.

The transfer films appearing after the very first contact cycles of linear polyolefins may significantly differ from the original polymer in response to the test conditions (load, velocity, temperature, surface roughness, etc.) The weight average molecular mass M_w of the transfer film polymer may decrease from twofold in case of HDPE (Belyi and Nevzorov 1985) to tenfold for UHMWPE (Ruben et al. 1993) compared with the original polymer; in case of PTFE, the estimates of mean number average molecular weight M_n of the debris polymer indicate an even deeper degradation, especially for the PTFE-based wear-resistant composites (Arkles and Schireson 1976). Tribo-degradation of PTFE involves liberation of fluorine and formation of fluorides on the metal counterbody surface; the ESCA spectra of the transfer films of PTFE, PA6, and PMMA show new, unidentified carbon-containing compounds (Jintang 2000). Orientation of the polymer macromolecules in the transfer film and in the polymer sample friction surface reduces friction from the initially higher level of static friction after several traversals at unidirectional or reciprocal friction tests and preserves the low friction value after renewal of sliding friction (Makinson and Tabor 1964; Schönherr and Vancso 1998). If sliding direction is permanently varied, friction remains at the initial high level of static friction.

The viscoelastic state of the amorphous interphase in PTFE crystallites determines high dependence of its friction and wear on test velocity, pressure, and temperature.

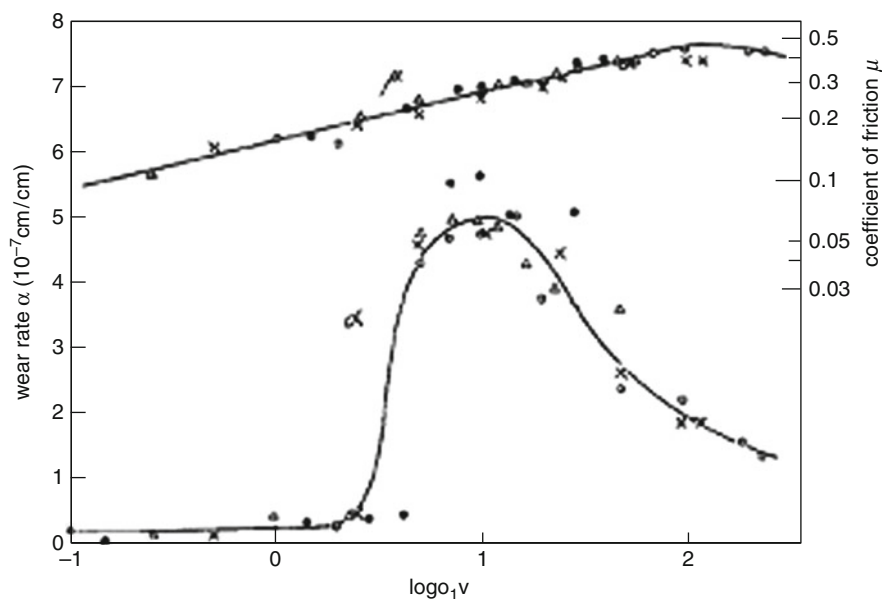
As sliding velocity exceeds 0.01 m/s, the friction coefficient increases from the low value of 0.03–0.1 typical of the low friction regimes to 0.3–0.4. Friction also increases to a smaller extent with lowering temperature (the sub-ambient temperature range) or reducing load. Increased friction at $V > 0.01$ m/s is accompanied with a two-order increase in the wear rate from $\sim 10^{-5}$ to $\sim 10^{-3}$ mm³/(N × m), or the onset of catastrophic wear (Fig. 1). Combined with high creep, it significantly limits application of the neat PTFE in dry friction units and necessitates its filling, an efficient way of strengthening and increasing the wear resistance of the material.

HDPE melting temperature is about 200°C lower than that of PTFE. It is also weaker in terms of thermooxidative and heat resistance, which limits its application in either pure or filled form. Increase of the M_w of HDPE over 10^6 modifies its mechanical properties, increasing toughness and impact strength, especially at cryogenic temperatures. Wear resistance is also significantly improved (Table 1).

The drawbacks of pure PTFE and UHMWPE like high creep and low heat resistance, are aggravated in the case of PTFE, with catastrophic wear at sliding speeds above 0.01 m/s, and are overcome either by filling or cross-linking. It should be stressed, that wear resistance of the linear polyolefins is highly dependent on counterface roughness, e.g., the wear rate of UHMWPE decreases about three orders of magnitude as the counterface roughness is reduced from $R_a = 0.3\text{--}0.5$ μm to an optimal roughness $R_a = 0.14\text{--}0.24$ μm at sliding speeds of 1–5 m/s (Barrett et al. 1992). Wear resistance of γ -cross-linked HDPE rises by 2 orders of magnitude (Pleskachevskii et al. 1981), that of g-crosslinked UHMWPE by 1–2 orders of magnitude (Oonishi et al. 1992), and that of γ -cross-linked at $T > T_m$ PTFE by 4 orders of magnitude (Khatipov and Artamonov 2008) to K_w reaching $\sim 1 \times 10^{-8}$ mm³/Nm.

Self-lubricating Materials Based on Linear Polyolefins

Due to high intrinsic lubricity of linear polyolefins, the fillers used may not possess lubricating properties, especially in case of PTFE-based SLM, which may contain large proportions of stainless steel, nickel, or titan powders, unacceptable in formulations of SLM on other polymer matrices. Various fillers efficiently reduce PTFE wear rate and its dependence on sliding velocity and temperature, especially under the friction conditions causing catastrophic wear, allowing to reduce the wear rate by 2–4 orders of magnitude compared with that of neat PTFE. Filling usually raises friction at small speeds, but at $V > 0.01$ m/s the difference is negligible, except for the



Solid Lubricants, Polymer-Based Self-Lubricating Materials, Fig. 1 Master curve dependences of friction coefficient μ and wear rate α on sliding velocity combining the corresponding dependences at test temperatures of 23°C - ○, 50°C - ×, 70 °C - △, 100 °C - ●. a_T -shift factor (Tanaka et al. 1973)

Solid Lubricants, Polymer-Based Self-Lubricating Materials, Table 1 Some properties of linear polyolefins

Property	PTFE	HDPE	UHMWPE
1. Density g/cm ³ 25°C, t/cm ³	2.15–2.24	0.96	0.93
2. Melting temperature, °C	327	132	137–143
3. Heat conductance, Wt/(m·K)	0.252	0.42–0.44	0.41
4. Coefficient of linear thermal expansion, °C ⁻¹ at ambient conditions	25×10^{-5}	10	20×10^{-5}
5. Service temperature range, °C	–200 + 260	–60 to +60	–120 to +100
6. Tensile strength, MPa	14–35	24–42	19–35
7. Flexural strength, MPa	11–14	–	40
8. Flexural modulus, MPa			
at 20°C	470–850	550–800	690
at –60°C	1,300–2,800	–	800 (at –40°C)
9. Sliding friction coefficient			
at $V \leq 0.1$ m/s	0.03–0.1	0.08–0.18	0.05–0.16
at $V > 0.1$ m/s	0.2–0.4	0.2–0.35	0.22–0.35
10. Degree of crystallinity, %	~50	60–70	~50
11. Adhesive wear rate, Kw, mm ³ /Nm	1×10^{-3} to 1×10^{-5}	1×10^{-6}	1×10^{-6} to 1×10^{-9}

composites filled with large ($>50 \mu\text{m}$), hard particles, which show $\mu \sim 0.35$ –0.4 and significant abrasive properties. Minimum abrasive properties combined with high wear resistance is observed for PTFE composites

filled with soft fillers, e.g., powders or fibers of PEEK (polyetheretherketone), PPS (polyphenylenesulfide), PI (polyimide), Kevlar (polyaramide fiber), POD (polyoxadiazole), and other thermally resistant polymers.

Solid Lubricants, Polymer-Based Self-Lubricating Materials, Table 2 Mechanical and tribological characteristics of several commercial PTFE composites

Property	Filler type and content, mass%					
	CF, 10%	CF 10%, carbon 10%	GF 15%	GF 15%, MoS ₂ 5%	Carbon 29%, graphite 3%	Bronze 60%
Tensile strength, MPa	21.2	–	25.5	18.8	7.0	17.0
Elongation at break, %	245	–	285	270	35	250
Hardness, shore D	–	–	62	–	70	70
^a μ	0.42	0.40	0.39	0.41	0.6	0.2
^a $K_w \times 10^{-6} \text{ mm}^3/\text{Nm}$	1.5	1.0	1.4	0.8	1.3	0.4

^aAt 2 m/s and P = 2 MPa**Solid Lubricants, Polymer-Based Self-Lubricating Materials, Table 3** Characteristics of UHMWPE and its composites (GUR 412/212)

Characteristic	Filler			
	None	Glass spheres, 30% ^a	Wood flour, 30% ^a	CaCO ₃ 30% ^a
Density, g/cm ³	936	1,140	1,040	1,080
Tensile yield strength, MPa	22	20	–	18
Tensile strength, MPa	32	28	16	26
Relative elongation at break, %	427	400	–	315
Charpy impact strength, kJ/m ²	140	100	21	67
Brinell hardness, MPa	40	47	45	44
Wear rate, arbitrary units	100	160	154	130
Flexural creep modulus after 7 days at 23°C and 4 MPa specific load, MPa	310	665	550	410

^aBy mass

Properties of some commercial PTFE-based tribological composites are given in Table 2 (Ebnesajjad 2000, http://www.fluon.jp/fluon/english/products/ptfe/grade_05.shtml).

Efficient wear-reducing fillers of HDPE (Briscoe et al. 1974) and some other polymers are sulfides, oxides, and fluorides of copper, lead, and some other metals, which is attributed to their positive effect on the rate of transfer film formation and its adhesion to the counterface. The fillers used to reinforce UHMWPE include various substances CaCO₃ (calcium carbonate), glass spheres, wood flour, etc. Such fillers are normally added from 10% to 30% by mass, and nanosized fillers from 0.1% to 2%. Similar to other polymers, addition of micro-sized fillers reduces impact strength, elongation at break, and tensile strength (Table 3).

Self-lubricating Materials Based on Other Polymers

Due to lower lubricity of the remaining polymer matrices, materials on their basis must contain one or several lubricants in addition to reinforcing fibers, normally introduced to raise the heat resistance and the loading capacity (PV limit), and special purpose additives, like antioxidants and UV stabilizers.

SLM may be classified based on the type of polymer matrix (thermoplastic or thermosetting), filler (fibrous, powder, continuous fiber, cloth, etc.), or application (highly loaded units, dry/lubricated, high/cryogenic temperatures, vacuum/other special medium, etc.).

Thermoplastic polymers are able to withstand melting-cooling cycle without significant change of

structure and properties. The group also includes thermoplastic elastomers, the mechanical behavior of which is similar to that of rarely cross-linked rubbers. Thermosetting polymers are produced by reaction of the reactive groups of oligomeric resin between themselves or with hardener molecules, yielding a non-melting polymer with a cross-linked structure. They are represented by phenolics, epoxies, unsaturated polyesters, cross-linked polyurethanes, organosilicon resins, cross-linked thermoplastics (PE, PTFE, etc.), and polyimides.

The reinforcements most widely used in SLM formulations include carbon fiber (CF), glass fiber (GF), aramide fibers (AF), cotton fiber, and their combinations. Much effort has been devoted to development of materials reinforced with natural fibers (sisal, jute, coconut) and layered silicates. Substitution of chopped GF with CF or AF reduces the wear rate up to an order of magnitude, also reducing the friction coefficient. The improvement of SLM lubricity with the addition of CF depends upon the

fiber type and orientation. Addition of lubricants reduces friction, which may raise the PV limit and often reduces the wear rate. Lubricants may be added either separately or in various combinations, e.g., PTFE is combined with silicon oils or CF, and graphite is combined with mineral oils, MoS₂, or PTFE. Nano-sized and submicron-sized fillers may additionally reduce friction and wear of pure polymers or fiber-reinforced SLM.

Tables 4 and 5 cite examples of several model and commercial SLM based on thermoplastic and thermosetting polymers to illustrate the effect of lubricants and reinforcing fibers, or fabrics in case of phenolic laminates (Table 5), on their performance under the conditions of adhesive wear. Laminated phenolics offer high load-carrying ability coupled with good wear resistance, both in dry and lubricated conditions, which makes them materials of choice for heavily loaded friction units. In that respect, they may compete with polymer-impregnated sintered porous metals, or bearings of impregnated

Solid Lubricants, Polymer-Based Self-Lubricating Materials, Table 4 Tribological characteristics of selected model SLM based on thermoplastic polymers

Polymer matrix	Filler	Lubricant	μ	$K_w, 10^{-6} \text{ mm}^3/(\text{N} \times \text{m})$
PA 46	–	–	0.8	20
PA 46		HDPE (20%)	0.25	2
PA 66	–	–	1.16	6.5
PA 66		CF (30%)	0.4	0.4
PA 66		CF(30%), 13% PTFE, 2% silicon oil	0.2	0.12
PA 66		Graphite, CF	0.35–0.7	0.5–5
PA 66	Nano-TiO ₂	Graphite, CF	0.26–0.44	0.5–3.3
PA 66	–	PTFE	0.13	0.5
PA 66	GF (30%)	PTFE	0.4	5.0
PA 66	AF	PTFE	0.3	0.22
PEEK	–	–	0.34–0.5	2–10
PEEK	–	CF (30%)	0.27–0.41	0.5–1
PEEK	–	PTFE (15%)	0.2	0.6
PEEK	GF (20%)	PTFE (15%)	0.2	1.1–1.3
PPS			0.6–0.8	40–70
PPS		PTFE (15%)	0.3	3
PPS	GF (20%)	PTFE (15%)	0.36–0.4	3.6
PPS		CF	0.55–0.6	0.5–0.7
PEI		–	0.4–0.45	30–45
PEI		CF (65%)	0.2–0.35	7–13
PEI		PTFE (15%)	0.16	0.6
PEI	GF (20%)	PTFE (15%)	0.2	2.0

Solid Lubricants, Polymer-Based Self-Lubricating Materials, Table 5 Tribological characteristics of selected SLM based on thermosetting polymers

Matrix polymer	Filler	Lubricant	PV, MPa m/s	μ	$K_w, 10^{-6} \text{ mm}^3 /$ (N × m)
Phenolformaldehyde polymer	Petroleum coke	–	–	0.25–0.3	2–4
	Chopped POD and cotton fibers	–	–	0.3	0.2–0.4
	Cotton fabric	–	0.6	0.1–0.3	–
	Cotton fabric	Graphite	1.1		–
	Cotton fabric	MoS ₂	1.8		–
	Cotton fabric	PTFE	2.5		–
	Glass fabric	–	0.9		–
	Glass fabric	Graphite	1.2		–
	Glass fabric	PTFE	2.5		–
Epoxy polymer	–	–	–	0.6	10–20
	CF-AF (44 vol%)	–	–		2.5
	CF-AF (47 vol%)	PTFE 5%	–		1.5
	–	–	2	0.6	1,000
	–	Nano-SiO ₂ (2 vol %)	2	0.3	10

sintered powdered bronze on steel backing. Elastomers are inferior to both thermoplastics and thermosets as matrices for self-lubricating materials because they show high friction losses and wear in dry friction units, so they are not considered here. Elastomer-based materials find tribological application mostly in lubricated friction units or are used as matrices of friction materials (see ► [Polymeric Elastomers: Tribological Behavior and Engineering Components](#), ► [Brake Friction Materials](#), ► [Clutch Friction Materials](#), ► [Polymeric Elastomers: Material Aspects of Tribology](#)).

Key Applications

PTFE finds wide use as a lubricant additive to polymer composites, coatings, greases, and oils. It is used in the form of fibers or fabrics, including hybrid fabrics of PTFE and CF, GF, or polymer fibers, also in the form of micropowders and waxes. Micropowders are produced by thermal, thermooxidative, or radiation-induced degradation of high molecular mass polymer ($M_w \sim 10^6$ – 10^7) in vacuum or controlled gaseous medium followed with mechanical milling to 3–20 μm (Ebnesajjad 2000). Degradation of the original PTFE powder in the presence of oxygen produces active $-\text{C}(\text{F})=\text{O}$ end-groups, improving adhesion of transfer films. Addition of PTFE reduces

friction and eliminates stick–slip, and increases the wear resistance and PV limit of SLM. PTFE content is varied between 5% and 30% by mass; the magnitude of the effect of PTFE addition depends on the dominant type of wear. The maximum effects are observed for adhesive and fretting wear: friction may be reduced two- or threefold, wear rate up to 10–20-fold, and loading capacity grows up to several times. In the case of abrasive wear, the tribological characteristics of the composite may deteriorate owing to reduced mechanical strength of SLM.

PTFE and PE polymer waxes are used as the mold release agents in processing of elastomers. UHMWPE, HDPE, and its degradation products are used, similar to PTFE, as additives to lubricating oils and greases, and less often as lubricants in SLM, e.g., in polyamide-based composites. PTFE, UHMWPE, and their composites are used in industry in seals, pumps, valves, and packings. The wide service temperature range of PTFE makes PTFE-based SLM invaluable for both cryogenic and high-temperature friction units. UHMWPE is also used to make gears, wear shoes/guides, sprockets, scraper blades, wear rails, star wheels, linings for chutes, hoppers, railcars, silos, bunkers, and dragline buckets. In the food industry it is used for cutting boards and wear components for food processing and handling equipment.

UHMWPE also serves to produce sports commodities, such as sled and ski liners, snowmobile parts, runways, artificial skating rinks, and amusement park slides. Biocompatibility combined with unsurpassed tribological properties make UHMWPE indispensable in arthroplasty of human joints, where it has been successfully used at an ever-increasing scale since the early 1960s.

Polymer-based SLM are used in various fields of mechanical engineering to reduce friction and wear losses of moving parts: seals, guides and linings in pneumatic and hydraulic cylinders, gear boxes, pumps, and compressors. In the automotive industry, SLM are used to make bushes, bearings, sliders, shock absorbers, ball bearing supports, door hinges, belt guides, window guides, chain tighteners, and rollers. Other applications include travelers in weaving and spinning machines, bearings in paper-making equipment, gears, dishwasher parts, conveyor components, and escalator guides.

Nomenclature

AF	Aramid fibers
CaCO ₃	Calcium carbonate
CF	Carbon fiber
GF	Glass fiber
HDPE	High-density polyethylene
Kevlar	Polyaramide fiber
MoS ₂	Molybdenum disulfide
PA	Polyamides
PE	Polyethylene
PEEK	Polyetheretherketone
PEI	Polyetherimide
PI	Polyimide
PMMA	Polymethylmethacrylate
POD	Polyoxadiazole
POM	Polyoxymethylene
PPS	Polyphenylenesulfide
PTFE	Polytetrafluoroethylene
SLM	Self-lubricating materials
UHMWPE	Ultrahigh molecular weight polyethylene

Cross-References

- [Brake Friction Materials](#)
- [Bonded Solid Lubrication Coatings, Process and Applications](#)
- [Clutch Friction Materials](#)
- [High-Temperature Solid Lubricating Materials](#)

- [MoS_x Coatings by Closed-Field Magnetron Sputtering](#)
- [Nanocomposite Coatings](#)
- [Polymer Composites and Nanocomposites](#)
- [Polymeric Elastomers: Material Aspects of Tribology](#)
- [Polymeric Elastomers: Tribological Behavior and Engineering Components](#)
- [Polymers for High Temperature Tribo-Applications](#)
- [Polymers for Low-Temperature Tribology Applications](#)
- [Self-lubricating Hard/Ultra-Hard Coatings](#)
- [Solid Lubricant Films Deposited by Burnishing](#)
- [Solid Lubricants](#)
- [Solid Lubricants for Space Mechanisms](#)
- [Solid Lubricants, Ceramic-based Self-Lubricating Materials](#)
- [Solid lubricants, Layered-Hexagonal Transition Metal Dichalcogenides](#)
- [Solid Lubrication in Fretting](#)
- [Solid Lubricants, Graphene](#)

References

- B.C. Arkles, M.J. Schireson, The molecular weight of PTFE wear debris. *Wear* **39**(1), 177–180 (1976)
- T.S. Barrett, G.W. Stachowiak, A.W. Batchelor, Effect of roughness and sliding speed on the wear friction of ultra-high molecular weight polyethylene. *Wear* **153**, 331–350 (1992)
- V.A. Belyi, V.V. Nevzorov, Molecular features of transfer fragments when high-density polyethylene rubbed against metals, in *Polymer Wear and Its Control*, ed. by L. Lee (ACS, Washington, DC, 1985), pp. 205–212
- B.J. Briscoe, A.K. Pogorian, D. Tabor, The friction and wear of high density polyethylene: the action of lead oxide and copper oxide fillers. *Wear* **27**(1), 19–34 (1974)
- M.E. Campbell, J.B. Loser, E. Sneegas, *Solid Lubricants, Technological Survey*, NASA SP-5059, (NASA, US Government Printing Office, Washington, DC, 20402, 1966)
- S. Ebnesaajad, *Fluoroplastics Volume 1: Non-Melt Processible Fluoroplastics. The Definitive User's Guide and Databook* (Plastics Design Library/William Andrew Inc., Norwich, NY, 2000)
- G. Jintang, Tribochemical effects in formation of polymer transfer film. *Wear* **245**(1–2), 100–106 (2000)
- S.A. Khatipov, N.A. Artamonov, Development of a novel antifrictional and sealing material based on radiation-modified polytetrafluoroethylene, (in Russian). *Russian Chem J* **52**(3), 89–97 (2008). *Journal of D.I. Mendeleyev Russian Chemical Society*
- K.R. Makinson, D. Tabor, The friction and transfer of polytetrafluoroethylene. *Proc. R. Soc. Lond. Ser. A* **281**, 49–61 (1964)
- H. Oonishi, Y. Takayama, E. Tsuji, Improvement of polyethylene by irradiation in artificial joints. *Radiat. Phys. Chem.* **39**(6), 495–504 (1992)
- Y.U.M. Pleskachevskii, V.V. Smirnov, S.V. Kopylov, E.B. Dubova, Study of frictional behavior of irradiated polyethylene. *Sov. J. Frict. Wear* **2**(6), 64–67 (1981). Allerton Press
- C.M. Pooley, D. Tabor, Friction and molecular structure: the behavior of some thermoplastics. *Proc. R. Soc. Lond. A.* **329**, 251–274 (1972)
- G.C. Ruben, T.A. Blanchet, E.E. Kennedy, Formation of UHMWPE polymeric transfer films on sliding glass counterfaces: early and

steady-state wear studied by transmission electron microscopy. *J. Mater. Sci.* **28**, 1045–1058 (1993)

H. Schönherr, G.J. Vancso, The mechanism of PTFE and PE friction deposition: a combined scanning electron and scanning force microscopy study on highly oriented polymeric sliders. *Polymer* **39**(23), 5705–5709 (1998)

K. Tanaka, Y. Uchiyama, S. Toyooka, The mechanism of wear of polytetrafluoroethylene. *Wear* **23**, 153–172 (1973)

Solid Lubrication Coating with Reversible Self-Adaptation

► Chameleon or Smart Solid Lubricating Coatings

Solid Lubrication in Fretting

D. B. LUO¹, V. FRIDRICI², PH. KAPSA²

¹School of Mechanical Engineering, Southwest Jiaotong University, Chengdu, People's Republic of China

²Laboratoire de Tribologie et Dynamique des Systèmes, UMR CNRS 5513, École Centrale de Lyon, Ecully, France

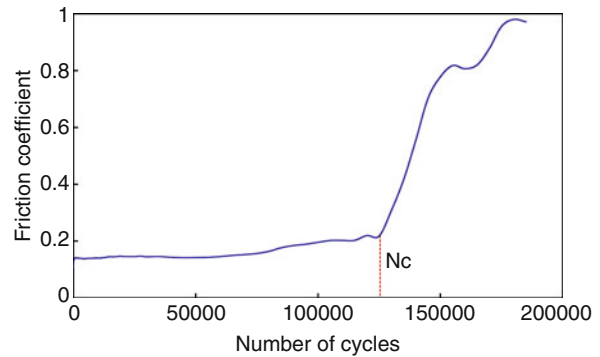
Definition

Fretting is a small-amplitude oscillatory motion between two contacting surfaces (due to vibrations, for instance), and it usually gives rise to wear or crack nucleation and propagation in tight assemblies where vibrations are likely to happen. Solid lubricants, generally in the coating form, can effectively reduce friction and palliate fretting damage.

Scientific Fundamentals

Effect of Solid Lubricant in Fretting

Fretting usually takes place in tight assemblies (such as suspension cables, dovetail contact in disk-blade system, electrical connectors, heat exchangers, and stem/bone contact in hip prostheses) where vibrations are likely to happen to induce small-amplitude oscillatory motion between contacting surfaces (Fridrici et al. 2003). According to the imposed displacement amplitude, normal load and material properties, three fretting regimes can be identified: partial slip regime, mixed slip regime, and gross slip regime. Fretting can result in wear (main damage in gross slip regime) or crack nucleation and propagation (main damage in mixed regime) (Zhou and Vincent 1995). Modification of the interface by

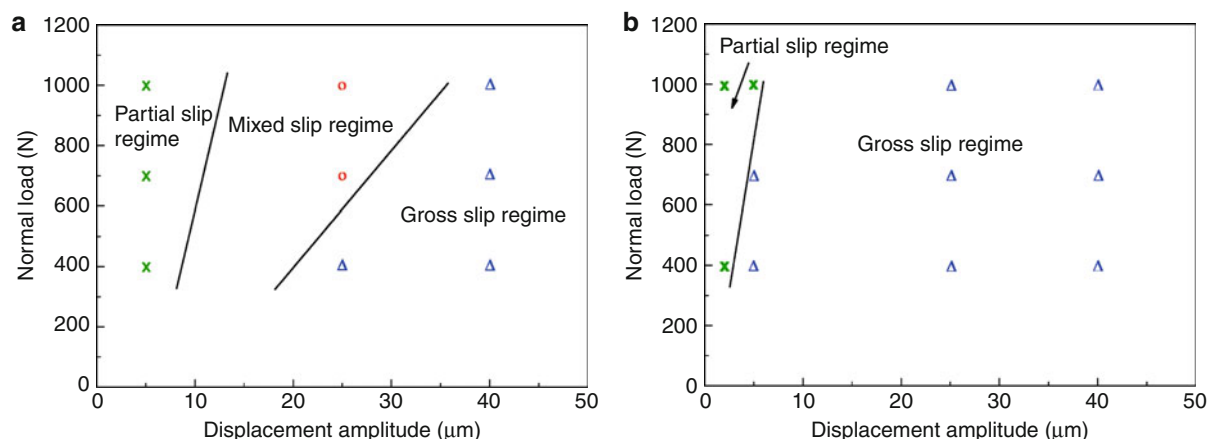


Solid Lubrication in Fretting, Fig. 1 Evolution of friction coefficient with number of fretting cycles for a bonded MoS₂ coating

introducing solid lubricant in the coating form to reduce the fretting contact loadings is an effective strategy to palliate fretting damage (Fouvry et al 2006).

Generally, the good friction reduction of solid lubricants results from the low shear strength at the interface, and with the degradation of lubricants, friction coefficient rapidly increases after the critical number of cycles (N_c) due to a direct contact of the substrate surface and the counterface (Fig. 1). Aside from the effect on friction, the introduction of solid lubricants also changes the fretting map. Figure 2a presents the running condition fretting map of an AISI 52100 cylinder rubbing against an SUS 316 block, where three fretting regimes can be identified. After the block is covered with a pressure-sprayed molybdenum disulfide (MoS₂) coating, under the same test conditions, the gross slip regime extends and the mixed slip regime disappears because the low friction of MoS₂ coating promotes the relative slip (Fig. 2b).

The durability of solid lubricant in fretting is strongly influenced by the preparing process. For a given technique, coatings of a solid lubricant prepared by different process parameters (substrate surface pretreatment, substrate temperature and coating thickness, etc.) present quite different wear lifetimes (Xu et al. 2003). The tribological performance of solid lubricant also depends on running conditions, like displacement amplitude, normal load, frequency, and environment. During a friction process, solid lubricant material can be cracked and detached due to changes of phase and structure or low adhesion strength. However, compared with sliding, in fretting, the detached material can easily be trapped in the contact area due to the low displacement amplitude. Instead of being ejected as debris, the trapped coating material still works at the contact area as third bodies. Under the cyclic



Solid Lubrication in Fretting, Fig. 2 Running condition fretting map of (a) ball on uncoated block and (b) ball on MoS₂ coated block

movements, the lubricity and durability of the coating depend on flows and rheology of third bodies, which are influenced by their cohesion, ductility, and running conditions (Descartes and Berthier 2002). With an increase of displacement amplitude, third bodies can be more easily thrown out from the contact area, and coating lifetime obviously decreases (Luo et al. 2010). An increase of normal load is also a cause for coating failure, because a high load can induce severe plastic deformation and detachment of particles (Xu et al. 2004).

Common Solid Lubricant in Fretting

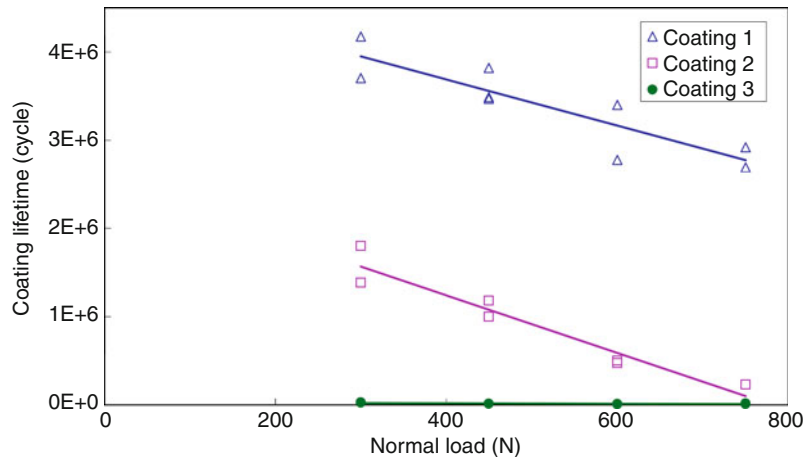
The common solid lubricants used in fretting include MoS₂, polymers, soft metals, and some self-lubricating composites.

MoS₂ is the common solid lubricant used in fretting due to its unique lamellar hexagonal structure. Bonding is a widely used technique for MoS₂ coating because it is a simple and cheap process. The adhesion strength and oxidation process restrict the employment of bonded MoS₂ coating in fretting (Zhu and Zhou 2001). During the service lifetime, the friction coefficient increases slightly with test process due to the gradually oxidation of MoS₂ (Fig. 1). The presence of water vapor can promote failure, because the most likely oxidation process is $\text{MoS}_2 + 2\text{H}_2\text{O} \rightarrow \text{MoO}_2 + 2\text{H}_2\text{S}$, then $\text{MoO}_2 + 1/2\text{O}_2 \rightarrow \text{MoO}_3$ (Fridrici 2002). Therefore, MoS₂ coatings are generally used for space applications. Because of advances in PVD technique, some MoS₂/metal composite coatings and multilayer coatings have been developed, which increases the possibility of MoS₂ coatings being employed in terrestrial atmosphere (Teer et al. 1997). For example, MoS₂/metal composite coatings deposited by a closed-field

unbalanced magnetron sputter ion plating technique are superior to pure MoS₂ coating in adhesion strength and tribological properties; MoS₂/Ti coating especially presents a lower humidity sensibility than MoS₂ (Teer et al. 1997). However, under fretting conditions, a better tribological performance of MoS₂/Ti coating than MoS₂ coating is not found, because the low strength of MoS₂ probably facilitates the formation of transfer film (Zhu et al. 2003).

Polymers, especially polytetrafluoroethylene (PTFE), are another kind of common solid lubricant in fretting. The tribological behaviors of bonded PTFE coatings have been widely investigated (Xu et al. 2004; Luo et al. 2010). The durability of bonded PTFE coatings strongly depends on the deposition process. For example, Fig. 3 presents fretting lifetime of three bonded PTFE coatings, which were prepared with different bonding agents and different cure processes. Coating 1 showed a very good durability, while the coating 3 rapidly degraded with severe detachments due to its adhesion problem. Under the same fretting conditions, PTFE coatings show lower and more stable friction coefficient than MoS₂ coatings. If there is no adhesion problem, the former also presents better durability than the latter (Luo et al. 2010). Around the contact area, strip debris and powder debris can be found, respectively, for PTFE coatings and MoS₂ coatings, which reveals that PTFE coatings have better ductility and the transfer film of PTFE can maintain a longer time to break. Therefore, the better tribological performance of PTFE coatings can be explained by the good ductility and relative inertness with atmospheric environment.

Soft metals (such as In, Sn, Cu, Au, Ag, and Pt) with a low shear strength are often deposited by electroplating



Solid Lubrication in Fretting, Fig. 3 Fretting lifetime of three bonded PTFE coatings under the same running conditions

or PVD as low-friction films. Because of the excellent electrical and thermal conductivity, oxidation resistance, and a relatively high melting point, soft metal films are usually used in electronic and electrical connections, where fretting can give rise to a rapid increase of contact resistance due to the accumulation of wear debris and oxides in the contact zone. A film of noble metals (Au, Pd, and their alloys) or non-noble metals (Sn, Sn–Pb) on copper alloys can minimize the potential for corrosion and improve durability (Park et al. 2007).

Diamond-like carbon (DLC) coating is a common protective coating in tribological applications due to the combination of high hardness, chemical inertness, low friction, and excellent wear resistance. In gross slip fretting tests, a-C:H coating presents a significant tribological performance, with a friction coefficient in the range of 0.01–0.1. The low friction mainly results from the formation of carbon transfer film on the counterface. The tribological behaviors of a-C:H are strongly influenced by test conditions, like counterpart, liquid lubrication, relative humidity, and temperature. The effect of counterpart on friction coefficient is very notable. When rubbing against an Al_2O_3 ball, the friction coefficient is low and stable, almost independent of relative humidity. When rubbing against a Si-based ceramic (like SiC or Si_3N_4) ball, the friction coefficient is high and fluctuating under a low relative humidity. For most counterpart materials, the wear resistance of a-C:H coatings increases with a decrease of relative humidity. Friction coefficient first decreases with an increase of temperature (the lowest value at 100°C), and then it increases, while the wear rate of a-C:H always increases with the increase of temperature (Wäsche and Klaffke 2008).

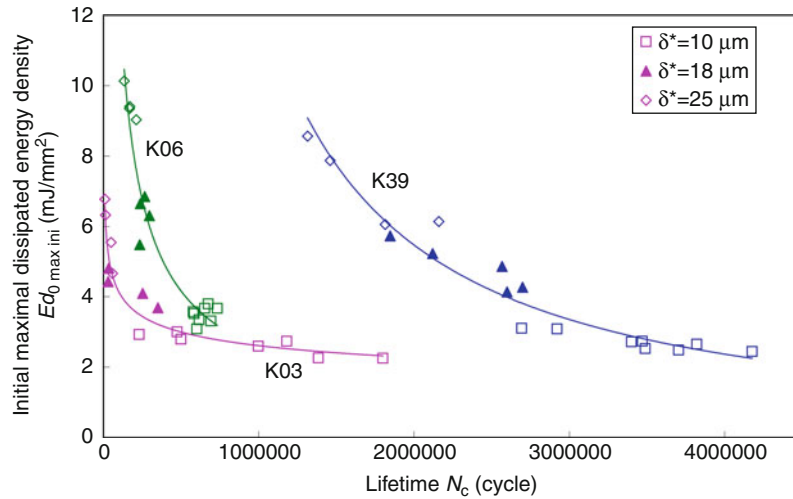
Selection of Solid Lubricants

With the development of coating deposition techniques, there are more and more available solid lubricant coatings. Only a reasonable selection strategy can result in a good tribological performance and a long coating lifetime. For a specific fretting application, how does one choose the optimal coating from the many possibilities?

Firstly, for a given application, there are some requirements and limits for the properties of the solid lubricant and for the coating deposition process. Therefore, some solid lubricant coatings, which do not suit the situation of the application, have to be excluded. For example, in a moist environment, MoS_2 is not a good choice, while graphite can provide a very low friction coefficient. MoS_2 presents its best tribological performance at around 100°C, and the highest temperature for PTFE is about 250°C, so these cannot be selected for high temperature applications. Other non-tribological performance requirements should also be considered.

Secondly, some poor solid lubricant coatings may be screened out by simple evaluation techniques. Mechanical properties (like hardness, elastic modulus, adhesion strength, and ductility) can be important for the durability of coatings. Some simple techniques (like nanoindentation, scratch test, and ball cratering test) can rapidly evaluate mechanical properties and screen out poor coatings, and save time for further simulation experiments. For example, bonded solid lubricant coatings with a good ball cratering resistance correspond to a long fretting lifetime (Luo et al. 2010).

Thirdly, evaluate fretting performance by simulation tests. The tribology performance, especially the wear resistance, is the most important factor for the selection of



Solid Lubrication in Fretting, Fig. 4 $Ed_{0 \max ini}$ – Fretting lifetime master curves of three bonded solid lubricant coatings (Luo et al. 2010)

solid lubricant. However, the friction coefficient and durability of a solid lubricant coating are influenced by values of test parameters (like displacement amplitude and normal load), so different curves between coating lifetime and a test parameter can be fitted with a change of another parameter. It is difficult to evaluate and compare the tribological performance of coatings by changing curves, especially when superposition takes place among curves of different coatings. Dissipated energy unifies normal load, sliding distance, and friction coefficient as one parameter, so it is a promising approach for comparing coatings. In general, there is a linear relationship between wear volume and cumulated dissipated energy. However, for the durability of coatings, the important index is not wear volume but the maximal wear depth, which is related to the maximal cumulated dissipated energy density, i.e., the maximal local cumulated dissipated energy (Fouvry et al. 2006). It is difficult to obtain the maximal cumulated dissipated energy density, because the contact area and the maximal contact pressure evolve with the friction process. Therefore, a simple initial maximal dissipated energy density $Ed_{0 \max ini}$ approach is more practical (Eq. 1), where only the first few cycles are involved (Fridrici et al. 2003).

$$Ed_{0 \max ini} = 4 \cdot \mu_{ini} \cdot p_0 \cdot \delta_{0 ini} \quad (1)$$

where p_0 is the maximal contact pressure, μ_{ini} and $\delta_{0 ini}$ are, respectively, initial friction coefficient and initial actual displacement amplitude.

Based on test results of three bonded coatings (K03 and K39 based on PTFE, K06 based on MoS_2) under

gross-slip fretting conditions, the relationship between $Ed_{0 \max ini}$ and fretting lifetime can be fitted by one master curve for each coating regardless of the values of test parameters (Fig. 4). K39 presents the longest fretting lifetime, and K03 is better than K06 for a low value of $Ed_{0 \max ini}$, or else K06 is better. Therefore, evaluation for the fretting resistance of the coatings can be straightforwardly obtained according to the curves.

Finally, a polar diagram can be employed to comprehensively compare coatings. In this approach, fretting behavior of coatings is straightforwardly compared through parameters from four aspects: intrinsic properties, coating-substrate interaction, running conditions, and material response (Carton et al. 1995). In fact, the polar diagram can be flexibly designed, i.e., parameters should be selected according to the given fretting application, even including non-tribological and non-functional requirements (Luo et al. 2010).

Key Applications

The typical application of solid lubricants in fretting is a rough and soft Cu-Ni-In under layer and MoS_2 coating in the blade-disk connection of turbo engines. In the dovetail joint of the blade-disk connection, the contacting surfaces suffer vibrations with a frequency of 200–400 Hz and a tangential sliding amplitude of about 5 μm . Observations of the failure surfaces imply the typical characteristics of degradation caused by fretting. A plasma sprayed Cu-Ni-In coating is deposited to protect the Ti-6Al-4 V substrate due to its plastic accommodation and low tangential contact stiffness. The Cu-Ni-In coating cannot

reduce the friction and the wear depth, but it results in a soft and compliant interface between the contacting bodies, which decreases the fretting contact loadings. The Cu-Ni-In coating reduces the length of the cracks observed on the counterbody and slightly moves up by 5 μm the mixed slip and gross slip regimes (Fridrici 2002). A low-friction bonded MoS_2 coating covered on the Cu-Ni-In further decreases the fretting contact loadings. Therefore, the treatment: plasma sprayed Cu-Ni-In + bonded MoS_2 is an effective countermeasure to prevent fretting damage, which has been widely employed in the blade-disk connection of turbo engines. Now, some new treatments are under development, e.g., mechano-chemically deposition techniques are used to deposit Mo + S and C-DLC lubricant coatings against fretting wear in aerospace components, which present a good adhesion strength and a significant improvement of tribological characteristics (Korsunsky et al. 2008).

Cross-References

- Bonded Solid Lubrication Coatings, Process and Applications
- Diamond-Like Carbon Coatings
- MoS_x Coatings by Closed-Field Magnetron Sputtering
- Self-lubricating Hard/Ultra-Hard Coatings
- Solid Lubricants
- Solid Lubricants, Layered-Hexagonal Transition Metal Dichalcogenides
- Solid Lubricants, Polymer-Based Self-lubricating Materials

References

- J.-F. Carton et al., Basis of a coating choice methodology in fretting. *Wear* **185**, 47–57 (1995)
- S. Descartes, Y. Berthier, Rheology and flows of solid third bodies: background and application to an $\text{MoS}_{1.6}$ coating. *Wear* **252**(7–8), 546–556 (2002)
- S. Fouvry et al., Palliatives in fretting: a dynamical approach. *Tribol. Int.* **39**(10), 1005–1015 (2006)
- V. Fridrici, Fretting d'un alliage de titane revêtu et lubrifié: application au contact aube/disque, Ph. D. thesis, Ecole Centrale de Lyon, Ecully, 2002, p. 200
- V. Fridrici et al., Impact of contact size and geometry on the lifetime of a solid lubricant. *Wear* **255**(7–12), 875–882 (2003)
- A.M. Korsunsky et al., Development and characterization of low friction coatings for protection against fretting wear in aerospace components. *Thin Solid Films* **516**, 5690–5699 (2008)
- D.B. Luo et al., Selecting solid lubricant coatings under fretting conditions. *Wear* **256**(5–6), 816–827 (2010)
- Y.W. Park et al., Fretting corrosion of tin-plated contacts: evaluation of surface characteristics. *Tribol. Int.* **40**, 548–559 (2007)
- D.G. Teer et al., The tribological properties of MoS_2 /metal composite coatings deposited by closed field magnetron sputtering. *Surf. Coat. Technol.* **94–95**, 572–577 (1997)
- R. Wäsche, D. Klaffke, Tribology of DLC films under fretting conditions, in *Tribology of Diamond-Like Carbon Films: Fundamentals and Applications*, ed. by C. Donnet, A. Erdemir (Springer, New York, 2008)
- J. Xu et al., An investigation on fretting wear life of bonded MoS_2 solid lubricant coatings in complex conditions. *Wear* **255**(1–6), 253–258 (2003)
- J. Xu et al., Fretting wear behavior of PTFE-based bonded solid lubrication coatings. *Thin Solid Films* **457**(2), 320–325 (2004)
- Z.R. Zhou, L. Vincent, Mixed fretting regime. *Wear* **181–183**, 531–536 (1995)
- M.H. Zhu, Z.R. Zhou, An investigation of molybdenum disulfide bonded solid lubricant coatings in fretting conditions. *Surf. Coat. Technol.* **141**, 240–245 (2001)
- X. Zhu et al., Different tribological behavior of MoS_2 coatings under fretting and pin-on-disk conditions. *Surf. Coat. Technol.* **163–164**, 422–428 (2003)

Solid-Like Lubricating Films, Ionic Liquid Films

WENJIE ZHAO

Ningbo Key Laboratory of Marine Protection Materials,
Ningbo Institute of Material Technology and Engineering,
Chinese Academy of Sciences, Ningbo,
People's Republic of China

Definition

Ultra-thin ionic liquid (IL) films with a thickness of only a few nanometers were prepared by dip-coating, spin-coating, or self-assembly methods. IL molecules can form a solid-like protective boundary film by physically adsorbing or chemically reacting to the substrate surface. Ultra-thin IL films have proved to be effective and beneficial for anti-adhesion, friction-reduction, and anti-wear on several surfaces and demonstrate promising potential as lubricant films for various device applications, such as M/NEMS and magnetic recording systems.

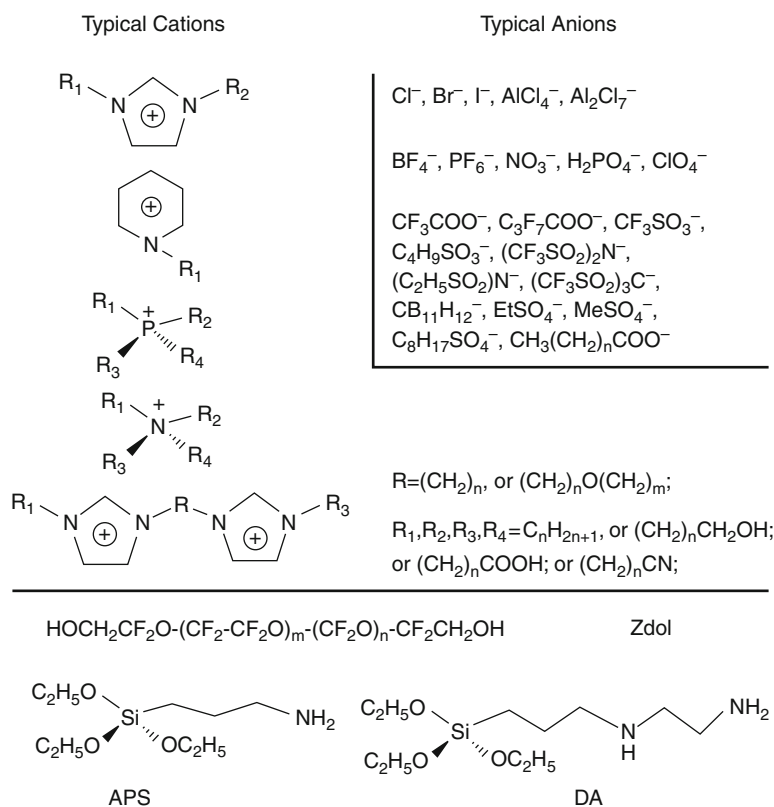
Scientific Fundamentals

Advances in nanotechnology have boosted the rapid development of device miniaturization, better integrated functional components, and energy-saving properties. Miniaturized devices such as M/NEMS and hard disk magnetic storage systems, have large surface-to-volume ratios and are operated under very small normal loads. Surface phenomena such as adhesion, friction, and wear are key issues and big challenges for the realization and reliability of many M/NEMS devices. It is well known that for M/NEMS the ideal lubricant should be molecularly thick, easily applied, able to chemically bond to the

micro/nano-device surface, insensitive to environment, and highly durable (Palacio and Bhushan 2010). Various ultra-thin films such as self-assembled monolayers (SAMs) are being developed and extensively studied. SAMs have been widely investigated due to the simple preparation process, molecularly thin feature, strong chemical bonded interface, high packing density, and robust stability. More importantly, it is possible to control the physical and chemical properties and desired chemical functionality with a unique molecular architecture. Using SAMs the surface structure can be controlled at the atomic scale by modifying the terminal group of the SAM molecule. However, the durability of SAMs is not good because there are no tribological contributions of the mobile molecules, such as replenishment, low shear strength, trapping among the micro-asperities, and so on. Also, their reliability and feasibility still need to be improved by combining with other lubricant materials.

ILs composed of bulky, asymmetric organic cations and weakly coordinating anions are molten salts at relatively low temperature. ILs possess a combination of unusual characteristics, including negligible vapor

pressure ("green" lubricants), non-flammability, high thermal conductivity (which helps to dissipate heat during friction) and stability (the decomposition temperatures of imidazolium ILs are generally above 350°C, some even as high as 480°C), high polarity (the most notable characteristic that distinguishes ILs from other synthetic lubrication oils), low melting point, excellent electrical conductivity (electrical conducting lubricants are needed for various electrical applications), miscibility with water and organic solvents, the number of combinations of anions and cations of ILs is in the range of one million, low cost compared to conventional synthetic organic lubricants (providing further motivation for evaluating the suitability of ILs as lubricants) (Palacio and Bhushan 2010; Zhou et al. 2009). These unique advantages are just what high performance lubricants demand, so ILs are being considered for M/NEMS applications because of their above-mentioned unique characteristics and superior lubrication performances. Molecular structures of common ILs cations and anions, Zdol, two kinds of SAMs, are shown in Fig. 1. As shown in Fig. 1, imidazolium, pyridinium, phosphonium, and



Solid-Like Lubricating Films, Ionic Liquid Films, Fig. 1 Chemical structures of common ILs, Zdol, and two kinds of SAMs

ammonium are the most frequently used cations, R stands for an organic group. It should also be noted that the organic substituents are usually different from each other, which further increases the potential number of available IL cations. Chemical and physical properties of ILs can be tuned through careful selection of cation and anion.

Ultra-thin IL films have been proved to be effective and to benefit anti-adhesion, friction-reduction, and anti-wear on several surfaces and demonstrate promising potential as lubricant films for M/NEMS. Several kinds of IL films were exploited by a number of researchers aiming to understand the micro/nano-tribological behaviors in molecular details. It was observed that the adhesion and micro/nano-tribological behaviors of ultra-thin IL films were strongly dependent on three key factors including the molecular structures (including the anion, cation, and the functional group), bonding fraction (bonding, partially bonding, and unbonding), and working environment (relative humidity and temperature) (Palacio and Bhushan 2009, 2010; Zhao et al. 2009a, b, 2010, 2011; Zhu et al. 2009; Mo et al. 2008; Bhushan et al. 2008).

Molecular Structure (Anion, Cation, and Functional Terminal Groups)

Anion

Four kinds of ILs, including 1-butyl-3-methylimidazolium hexafluorophosphate (BMIM-PF_6^-), 1-butyl-3-methylimidazolium tetrafluoroborate (BMIM-BF_4^-), 1-butyl-3-methylimidazolium perchlorate (BMIM-ClO_4^-), and 1-butyl-3-methylimidazolium nitrate (BMIM-NO_3^-) were chosen to investigate the effect of anion on the micro/nano-tribological properties of IL films, Zdol was used as a comparison (Zhao et al. 2009a).

It was found that the adhesive force which was closely related to the anion of the IL films increased in the sequence of Zdol, BMIM-PF_6^- , BMIM-BF_4^- , BMIM-ClO_4^- and BMIM-NO_3^- . It is well known that, when the lubricant films were disordered and hydrophilic, they would easily form meniscus by themselves or the adsorbed water molecules, thus they showed higher adhesive force. However, when the lubricant films were hydrophobic and ordered, they would show low adhesion. Zdol film was hydrophobic and its surface energy was lower than other four kinds of IL films, it tended to form densely packed, highly ordered film, so it showed the lowest adhesive force. BMIM-BF_4^- , BMIM-ClO_4^- and BMIM-NO_3^- can dissolve in water, the water molecules may be easily adsorbed onto the surface of films from the ambient environment, so these films show larger adhesive force. Compared with the above

lubricants, BMIM-PF_6^- cannot dissolve in water, and also its film formed on the silicon surface may be not as ordered as Zdol film, so its adhesive force is between the above two.

The nano-friction force of the film increases from Zdol, BMIM-PF_6^- , BMIM-NO_3^- , BMIM-ClO_4^- , to BMIM-BF_4^- . The difference in the nano-friction of the films might be attributed to three potential factors: (1) intra-molecular energetic barriers to rotation of the rigid ring structure; (2) long-range intermolecular steric interactions within the plane of the bulkier groups; and (3) surface energy. All the ILs contained rigid ring-shape structure and they need much energy than Zdol to overcome intra-molecular energetic barriers to rotate the rigid cycle structure. And also if the surface energy is much larger, it is easily to form meniscus by themselves or the adsorbed water molecules, they had larger adhesive force due to the capillary force and hydrogen bond, which would led to larger shearing strength and higher friction. So all IL films exhibited higher friction force than Zdol, and the difference among the IL films was caused by diverse anions that exhibited different hydrophobic/hydrophilic properties. ILs also showed different micro-tribological performance due to different anions, some showed better tribological properties than Zdol, for example: BMIM-PF_6^- , BMIM-BF_4^- and BMIM-ClO_4^- ; but some showed worse micro-tribological performances than Zdol, such as BMIM-NO_3^- . BMIM-PF_6^- film was much superior to four other kinds of films. It can be concluded that the micro/nano-tribological performances were significantly determined by the anions of ILs.

The effect of anion on the micro/nano-tribological properties of several IL films on modified silicon wafers was also studied (Zhu et al. 2009). It was observed that in nanoscale, 1-hexyl-3-methyl-imidazolium hexafluorophosphate (IL106P), which owned the lowest surface energy, possessed the lowest adhesion force and friction coefficient, and bare aminated Si, which had lower surface energy performed lower adhesion and friction coefficient than bare hydroxylated Si. In microscale, 1-hexyl-3-methyl-imidazolium tetrafluoroborate (IL106B) and IL106P exhibited better tribological properties than 1-hexyl-3-methyl-imidazolium adipate (IL106A) due to their stable inorganic anions. The results and summary obtained by them agree well with the foregoing one.

Cation

Three kinds of IL films were evaluated, aiming to find the relationship between the micro/nano-tribological properties and cations of ILs. ILs including two dicationic

ILs 1, 1'-(pentane-1,5-diyl)bis(3-hydroxyethyl-1*H*-imidazolium-1-yl)di[bis(trifluoromethanesulfonyl)imide]] (abbreviated as BHPT) and 1,1'-(3,6,9,12,15-Pentaoxapentadecane-1,15-diyl)bis(3-hydroxyethyl-1*H*-imidazolium-1-yl)di[bis(trifluoromethanesulfonyl)imide]] (abbreviated as BHPET), and the common monocationic ionic liquid 1-butyl-3-methyl-imidazolium hexafluorophosphate, which abbreviated as BMIM-PF₆, was provided as a comparison (Palacio and Bhushan 2009).

BHPT exhibited superior nanoscale friction and wear resistance properties. BHPET showed less desirable adhesion, friction, and wear performances compared with either BHPT or BMIM-PF₆. The microscale friction coefficient was higher than the nanoscale value due to the difference in the length scales. BHPT was the most hydrophobic film among the three IL films and exhibited the largest reduction in friction coefficient relative to the uncoated Si surface. The low surface energy of BHPT led to minimal capillary force between the tip and surface, additionally, the presence of a pentyl chain and hydroxyl groups on both chain ends, which facilitated molecular orientation and bonding interactions with the substrate surface, leading to a large decrease in the nanoscale friction force. The BMIM-PF₆ film also exhibited a reduction in the adhesive force and friction coefficient due to the combination of mobile and immobile lubricant fractions. In contrast, the polyether chain of BHPET was susceptible to interactions with water molecules, which can promote meniscus formation. Intermolecular hydrogen bonding in BHPET also reduced the chain ordering on the substrate surface, which accounted for the observed larger adhesive force and friction coefficient relative to the other ILs investigated.

The influence of cation on the micro/nano-tribological properties of ultra-thin IL films was also investigated by choosing three kinds of ILs including 3-butyl-1-methyl-imidazolium tetrafluoroborate, tetraalkylphosphonium tetrafluoroborate and N-butylpyridinium tetrafluoroborate, marked as L-B104, I-P and I-N, respectively, and Zdol was also evaluated so as to have a comparison (Zhao et al. 2009b).

It was observed that L-B104 and I-N exhibited larger friction forces compared with Zdol in the nanoscale friction. The tribological behaviors of I-P and I-N were different at small loads, but the difference decreased with increasing load. The difference in nano-friction is attributed to three potential factors: firstly, intra-molecular energetic barriers; secondly long-range inter-molecular steric interactions; and finally surface energy. When the surface energy is much larger, it is easy to form a meniscus by themselves or the adsorbed water molecules, thus they

exhibited larger adhesive force due to the capillary force and hydrogen bond, which would lead to higher shearing strength and larger friction. Compared with the nanoscale friction, the micro-tribological properties were determined by the combination of soft chains and the rigid ring structure. The soft chains, which were flexible, can be used to decrease the friction and the rigid ring structure was used to resist the wear during the test. L-B104 contained both rigid ring structure and flexible chains, so it showed the best tribological performances among the lubricants used in this study. Zdol had free linear chains and it showed better tribological properties. I-P had many branched chains but without rigid ring structure, I-N had rigid ring structure but lack of flexible chain, so both of them showed poor micro-tribological behaviors.

Functional Terminal Groups (FTGs)

Different FTGs can apparently affect micro/nano-tribological performances of ultra-thin IL films. Four kinds of ILs with the same anion but different FTGs, such as 1-propyl-3-methylimidazolium chloride (MIMCH-CL), 1-ethanol-3-methylimidazolium chloride (MIMOH-CL), 1-propionitrile-3-methylimidazolium chloride (MIMCNCL), and 1-propionic acid-3-methylimidazolium chloride (MIMCOOH-CL) were evaluated in order to investigate the influence of FTGs on the micro/nano-tribological properties of ultrathin IL films (Mo et al. 2008).

The adhesive force was observed to decrease in the following order: Si>MIMOH-CL>MIMCOOH-CL>MIMCN-CL>MIMCH-CL. The largest adhesive force was observed on the hydroxylated Si surface and was decreased after the IL films were coated. It is well known that hydrophilic lubricant films easily form a meniscus by themselves or by the adsorbed water molecules, which results in a higher adhesive force due to the capillary effect. In turn, hydrophobic lubricant films show low adhesion. MIMCH-CL exhibits the lowest adhesive force since -CH₃ exhibited the most relative hydrophobic property among the four IL films investigated in this work. It was also observed that friction forces of MIMCOOH-CL and MIMOH-CL were larger than that for MIMCH-CL and MIMCN-CL, all the IL films greatly reduced friction force for hydroxylated Si surface. The results implied that the hydrophobic property of FTGs facilitated sliding of the spherical tip on the surface. Both of the FTGs of -COOH and -OH showed more hydrophilic than -CH₃ and -CN, water and lubricant molecules are more likely to cause a formation of capillary force as the spherical tip contacts with the surface. This provides greater resistance during tip sliding and leads to larger friction force.

All IL films exhibited great friction reduction and anti-wear durability in the microscale friction. MIMCN-CL and MIMCOOH-CL exhibited low friction coefficient and long anti-wear durability even at high normal load. The MIMOH-CL, MIMCN-CL, and MIMCOOH-CL exhibited lower friction and better anti-wear durability at high-frequency sliding compared with MIMCH-CL. It can be concluded that the optimum choice of the functional cations can greatly improve their micro-tribological properties.

The micro/nanotribological behaviors of four IL ultra-thin films (two with polar FTGs, such as $-\text{OH}$ and $-\text{COOH}$, and the other two with apolar FTGs, such as $-\text{CH}_3$ and $-\text{phenyl}$.) were also systematically investigated (Zhao et al. 2011). It is observed that adhesive force was determined by the hydrophobic/hydrophilic properties of ILs' FTGs. IL ultra-thin films with more polarized FTGs generally possessed higher surface energy and a relatively strong interaction occurred between the tip and the samples during the sliding, therefore, higher adhesion and more energy loss are expected. IL ultra-thin films with polar or stiff phenyl FTGs exhibited relatively larger friction forces and better anti-wear performances than the ones with apolar alkyl chain structure at micro/nanoscale. The different micro/nano-tribological performances of the IL ultra-thin films were mainly dependent on their different FTGs, which mainly influenced their hydrophobic/hydrophilic and soft/rigid chain structure properties.

Bonding Fraction (Bonding, Partially Bonding, and Unbonding)

Mixing films with the partially bonded and mobile lubricant exhibited the best lubrication characteristics. 1,3-di (2-hydroxyethyl)imidazolium hexafluorophosphate was carefully chosen to study the effect of bonding fraction on the micro/nano-tribological properties of IL films due to two beneficial effects of the unique properties of both cation and anion (Zhao et al. 2010). Firstly, the cation owns two OH-groups that can bond to the hydroxylated silicon surface. Secondly, the hexafluorophosphate (PF_6^-) anion showed better micro/nano-tribological performance than other anions. IL-OH films were heated at 60°C and 120°C for 1 h, respectively, and were designated as IL-OH-60 and IL-OH-120. IL-OH-120 film after being rinsed and washed ultrasonically with excess acetone to remove physisorbed IL-OH molecules was designated as IL-OH-120-clean.

The adhesive forces of the IL films changed in the sequence of $\text{Si-OH} > \text{IL-OH} > \text{IL-OH-60} > \text{IL-OH-120} > \text{IL-OH-120-clean}$. In other words, the adhesive force was observed to increase in the following order: fully bonded <partially bonded <untreated. The adhesive

forces were related to the bonding fraction of the molecules on the silicon surface, which was caused by the heat treatment. It is well known that disordered and hydrophilic lubricant films formed large capillary force easily by themselves or the adsorbed water molecules, resulting in larger adhesive force. Contrarily, hydrophobic and ordered lubricant films show low adhesion. IL films without being heat treated or treated at 60°C might not form highly uniform and fully bonded film; water is adsorbed easily to the film surface from the environment, hence showing larger adhesive force. Conversely, the sample with no mobile lubricant fraction exhibited the lowest adhesive force. IL-OH tended to form more uniform and fully bonded films after heat treatment. Accordingly, film with no mobile fraction (IL-OH-120-clean film) showed the lowest adhesive force.

Different bonding fraction of ILs also caused the different nano-tribological behaviors of IL films. At room temperature, the interaction between the lubricant and the substrate is weak, there is no (or almost no) bonded IL causing direct contact between tip and substrate, leading to larger friction force, so friction force for the untreated samples are larger than the heat treated coatings. IL-OH molecule binds stronger to silicon after heated at 60°C for 1 h, bonding and mobile fraction reach an appropriate degree. The mobile fraction present in the partially bonded film that can rotate with the tip sliding direction easily facilitates sliding of the tip on the surface. Therefore, IL-OH-60 film shows the lowest nano-friction force. After being heated at 120°C for 1 h, bonding fraction of IL-OH-120 film reaches the highest point and mobile fraction reaches the lowest point, hence, IL-OH-120 film shows larger nano-friction. IL-OH-120-clean film nearly has no mobile fraction, so it shows the largest nano-friction force among the tested IL films. In summary, with suitable heat treatment, bonding fraction and mobile fraction will reach an optimal proportion, leading to lower adhesion force and nano-friction force.

The microtribological results nicely correspond with the nanotribological performances. It was observed that the IL films with appropriate bonding fraction exhibited lower friction coefficient, longer durability, and better load-bearing capacity than larger bonding fraction and unbonded films in the test range of the load. The improved tribological performance of IL-OH-60 film should rely on its intrinsic structure. Both the mobile uplayer and rigid underlayer of IL-OH-60 films enhanced the stability and load-bearing capacity of the film. Briefly, designing IL film with good tribological performance should own the structure, which is composed of both a rigid part to withstand load and a mobile part to reduce

surface friction force. It is worth noting that the heat-treatment effect obtained in the investigation agrees well with other researchers who performed AFM-based nanotribological measurements on untreated and treated IL films on silicon (Bhushan et al. 2008).

As with the ultra-thin films on different substrates, one of the biggest challenges is the adhesion of ILs on substrates. In order to improve the interface of IL films with substrates, a lubrication scheme of synergy between bonded and mobile phases of ILs has been proposed to further improve their anti-rupture performance and durability by using SAMs as first layer, and the influence of different self-assembled underlayer on the micro/nanotribological properties of IL films with two-phase structure has been studied (Pu et al. 2010). Two kinds of dual-layer IL-COOH films that contain different self-assembled under-layers were formed on the silicon substrate. 3-aminopropyl triethoxysilane (APS) and N-[3-(trimethoxysilyl)propyl]ethylenediamine (DA) separately self-assembled on silicon surface as anchor layer at first, then IL molecules were chemically bonded to the silicon substrates modified by the SAMs to form the two-phase structure.

It was observed that DA-SAM, APS-SAM, and IL-COOH film showed larger adhesion forces than the dual-layer films, the DA-IL film showed the best anti-adhesion performance. Two kinds of SAMs exhibited very poor anti-wear capacity performance. As compared with single-layer IL-COOH film, the formation of chemically bonded phase in IL-COOH layer largely improved the micro/nano-tribological properties of the two kinds of dual-layer films that were attributed to synergic effect between flowing mobile phase and steady bonded phase. In particular, the DA-IL dual-layer film possessed excellent micro/nano-tribological properties, which were characterized by lower friction and higher anti-wear ability. The remarkable decrease of friction coefficient of dual-layer films might due to the combination of the mobile phase and bonded phase. There is no direct contact between the steel ball and silicon substrate during sliding because the underlying IL molecules chemically bonded on the APS/DA modified silicon substrate, which was hard to remove. Furthermore, a self-replenishment property was generated by the mobile IL molecules, which can flow back into the wear track due to their excellent mobility. Compared with APS-IL film, the DA-IL dual-layer film exhibited lower friction coefficient due to its more densely packed and ordered underlayer. DA with longer chains easily formed more densely packed and orderly SAM as compared with APS, which led to more densely packed bonded phase and meniscus reduction

effects. These characteristics of DA-IL led to its lowest friction coefficient among all the films. For the wear resistance durability, the improved wear resistance durability can be attributed to two potential factors: (1) the generation of the chemically bonded IL-COOH layer on silicon surface; (2) interlinked hydrogen-bonding among the molecules.

Working Environmental (Relative Humidity and Temperature)

The influence of the working environment including both relative humidity (RH) and temperature on the nanotribological behaviors was studied in detail (Palacio and Bhushan 2009, 2010). It was found that the adhesive force and friction coefficient increased with the RH in general. This was due to condensed water in the humid environment facilitated capillary force formation between the tip and film and thus showed larger adhesive forces. For all the IL films, attractive electrostatic interactions between the individual ions and water molecules led to increased water adsorption and also increased the adhesive force. The adsorbed water at higher humidity can lead to the formation of a continuous water layer separating the tip and sample surface, which can act as a lubricant. However, the presence of more water molecules showed an opposite effect on the IL surfaces at higher humidity, where the friction coefficient increased with humidity. More water molecules are available at higher humidity, which amplifies the attractive ion-dipole forces between the ions and water. A greater attractive force between the tip and surface led to larger resistance to sliding and a higher friction coefficient.

The effect of temperature on the nanotribological behaviors of the IL films that was measured from 22°C to 125°C is also investigated. The increase in test temperature led to a decrease in the adhesive force and the friction coefficient. The decrease in the adhesive force at higher temperatures is observed in all the films, but the corresponding drop in the friction coefficient was seen only for BMIM-PF₆, BHPET, and the silicon surface. When the temperature increased, the ability of water molecules absorbed on surface decreased significantly, leading to the decrease in both the adhesive and friction forces. A reduction in the viscosity at higher temperatures can also facilitate the decrease in the friction force.

Formation Mechanism of IL Films on Different Substrates

The favorable micro/nano-tribological behaviors are attributed to the polar nature of ILs, enabling them to physically or chemically adsorb or react to the substrate

surface, as well as form wear-resistant surface protective films during sliding. The immobilization of ILs can be carried out in several ways, which can be classified by the interaction between the ILs or their components and the substrate. ILs can be immobilized on a surface either by covalent bonds between silanol groups and the anion or the cation of the liquid, or without covalent bonds in the form of supported liquid phases (Valkenberg et al. 2002; Nooruddin et al. 2007, 2009).

The interactions between selected ILs and a modified silicon surface, an aluminum oxide, or a silicon nitride surface were modeled by using semi-empirical methods (Nooruddin et al. 2007, 2009). The modeled ILs included a series of ILs consisting of imidazolium derivatives with Cl^- as the anion differ only in the terminal group of a propyl side chain interacting with hydroxylated silicon wafers. A second series consists of symmetrical and asymmetrical dicationic imidazolium derivatives with PF_6^- or BF_4^- as the anion interacting with hydroxylated single crystal silicon wafers. Experimental results showed that there was an obvious correlation between friction coefficients and the enthalpies of various IL–silicon surface complexes, similar to that obtained for IL–silicon nitride and IL–Al–steel alloys.

Understanding the molecular interactions between ILs and several substrates by using semi-empirical methods has been proved to be useful and time-saving in the development of ILs that have excellent tribological properties and helps us to select optimal ILs for specific substrates.

Main Issues Concerning the Use of ILs as Lubricants

The main issues concerning the use of ILs as lubricants include IL causing corrosion on the substrate surface (especially ILs with halogen anions), thermal oxidation, and environmental toxicity, as well as the degradation by-products (Palacio and Bhushan 2010; Zhou et al. 2009). To date, corrosion issues can be resolved or minimized through two approaches: one is through the careful selection of cations and anions (replacing anions (primarily BF_4^- and PF_6^-) with more hydrophobic anions or with fluorine-free anions); another corrosion reduction strategy is the incorporation of anti-corrosion additives such as benzotriazole. Environmental toxicity of ILs is largely unknown. Research on micro/nano-tribological properties is also much diffused due to the diversity of available ILs. Therefore, all these problems need to be studied further in the future.

Key Applications

Ultra-thin IL films have been explored as protective films for various device applications including M/NEMS and magnetic recording system by worldwide scientists due to their unique chemical and physical properties. Recent investigations have shown that some ultra-thin IL films can match or even exceed the micro/nano-tribological behaviors of common high-performance lubricants such as PFPEs and X-1P. But study on ultra-thin IL films is just beginning; there are still many problems that need to be resolved in the future. To date, there are still no actual applications of ultra-thin IL films known. Further investigations of the molecular mechanisms that account for the micro/nano-tribological properties of ultra-thin IL films are necessary to systematically identify the optimal IL films for various sliding pairs under different working conditions. Ultra-thin IL films have proved to be effective and beneficial for anti-adhesion, friction--reduction, and anti-wear on several surfaces and demonstrate promising potential as lubricant films for M/NEMS. It is believed that ultra-thin IL films will play important roles in M/NEMS devices and magnetic recording system in the next few years.

Cross-References

- [Anti-adhesion/Stiction Surface Design, Fabrication, and Applications](#)
- [Bonding at Surfaces/Interfaces](#)
- [Capillary Force and Surface Wettability](#)
- [Friction Force Microscopy](#)
- [Ionic Liquid Lubricants](#)
- [Surface Analysis Using Contact Mode AFM](#)
- [Surface Forces, Surface Tension, and Adhesion](#)
- [Surface Free Energy](#)
- [Thin Film Lubrication](#)

References

- B. Bhushan, M. Palacio, B. Kinzig, AFM-based nanotribological and electrical characterization of ultrathin wear-resistant ionic liquid films. *J. Colloid Interface Sci.* **317**, 275–287 (2008)
- Y.F. Mo, W.J. Zhao, M. Zhu, M.W. Bai, Nano/microtribological properties of ultrathin functionalized imidazolium wear-resistant ionic liquid films on single crystal silicon. *Tribol. Lett.* **32**, 143–151 (2008)
- N.S. Nooruddin, P.G. Wahlbeck, W.R. Carper, Molecular modeling of IL tribology: semi-empirical bonding and molecular structure. *J. Mol. Struct.* **822**, 1–7 (2007)
- N.S. Nooruddin, P.G. Wahlbeck, W.R. Carper, Semi-empirical molecular modeling of ionic liquid tribology: ionic liquid–hydroxylated silicon surface interactions. *Tribol. Lett.* **36**, 147–156 (2009)
- M. Palacio, B. Bhushan, Molecularly thick dicationic ionic liquid films for nanolubrication. *J. Vac. Sci. Technol. A* **27**, 986–995 (2009)
- M. Palacio, B. Bhushan, A review of ionic liquids for green molecular lubrication in nanotechnology. *Tribol. Lett.* **40**, 247–268 (2010)

- J.B. Pu, D.M. Huang, L.P. Wang, Q.J. Xue, Tribology study of dual-layer ultrathin ionic liquid films with bonded phase: influences of the self-assembled underlayer. *Colloids Surf. A* **372**, 155–164 (2010)
- M.H. Valkenberg, C. DeCastro, W.F. Hölderich, Immobilisation of ionic liquids on solid supports. *Green Chem.* **4**, 88–93 (2002)
- W.J. Zhao, M. Zhu, Y.F. Mo, M.W. Bai, Effect of anion on micro/nano-tribological properties of ultra-thin imidazolium ionic liquid films on silicon wafer. *Colloids Surf. A* **332**, 78–83 (2009a)
- W.J. Zhao, Y.F. Mo, J.B. Pu, M.W. Bai, Effect of cation on micro/nanotribological properties of ultra-thin ionic liquid films. *Tribol. Int.* **42**, 828–835 (2009b)
- W.J. Zhao, Y. Wang, L.P. Wang, M.W. Bai, Q.J. Xue, Influence of heat treatment on the micro/nano-tribological properties of ultra-thin ionic liquid films on silicon. *Colloids Surf. A* **361**, 118–125 (2010)
- W.J. Zhao, L.P. Wang, Q.J. Xue, Micro/nanotribological behaviors of ionic liquid nanofilms with different functional cations, *Surf. Interface Anal.* **43**, 945–953 (2011)
- F. Zhou, Y.M. Liang, W.M. Liu, Ionic liquid lubricants: designed chemistry for engineering applications. *Chem. Soc. Rev.* **38**, 2590–2599 (2009)
- M. Zhu, Y.F. Mo, W.J. Zhao, M.W. Bai, Micro/macrotribological properties of several nano-scale ionic liquid films on modified silicon wafers. *Surf. Interface Anal.* **41**, 205–210 (2009)

Solid-Like Lubricating Films, Self-Assembled Films

JINQING WANG, JUNFEI OU, SHENGRONG YANG

State Key Laboratory of Solid Lubrication, Lanzhou
Institute of Chemical Physics, Chinese Academy of
Sciences (CAS), Lanzhou, People's Republic of China

Definition

As a new lubrication regime, thin film lubrication has been well studied since the 1990s. Lubrication with ordered films is a common practice in many modern technical devices. Generally, ordered films can be formed on surfaces (substrates) by several approaches, including the Langmuir–Blodgett (LB) method, self-assembly, adsorption, and evaporation. Among these, the self-assembly technique has generated substantial interest not only for its simple preparation procedure but also for its wide potential applications in many fields, such as surface modification, boundary lubricant coatings, sensors, photoelectronics, and functional bio-membrane modeling.

Self-assembled films (SAFs) are ordered molecular assemblies formed on a solid surface through a spontaneous process driven by certain forces. SAFs can be divided into two categories based on the number of layers in

the film: self-assembled monolayers (SAMs) and self-assembled multilayer films (SAMFs).

Scientific Fundamentals

Self-Assembled Monolayers (SAMs)

SAMs are molecular assemblies that are formed spontaneously on a solid surface by the immersion of a substrate into a solution containing active surfactant (Fig. 1a, Ulman 1996). A self-assembling precursor surfactant molecule includes three parts: a head group, an alkyl chain, and a tail group (Fig. 1b). Each part has great influence on certain properties of SAMs (Tsukruk 2001). Briefly, the head groups adhere to the substrate through certain strong chemical interactions and play an important role in determining the affinity of SAMs. The alkyl chains are packed together through interchain van der Waal forces and influence the pack density of the assemblies. The tail groups are exposed to the environmental atmosphere and serve as the surface layer, determining surface properties such as wettability, reactivity, and so on.

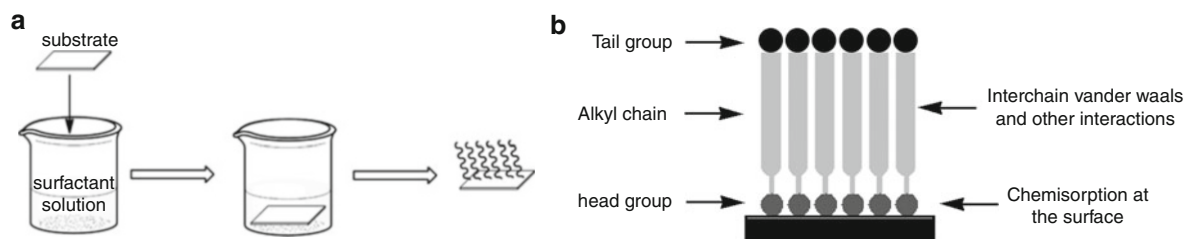
According to the interaction between head group and the substrate, SAMs can mainly be divided into the following three categories.

Organosilane Monolayers

SAMs of organosilanes, such as alkylchlorosilanes, alkylalkoxysilanes, and alkylaminosilanes, are generally formed on the hydroxylated surface of oxide (MO_x), such as SiO_2 and Al_2O_3 . The driving force for this self-assembly is the in situ formation of Si–O–M interfacial bonding. Taking the assembly of *n*-octadecyltrichlorosilane (OTS) as an example, it is supposed that the OTS molecules would first be hydrolyzed under the attack of the adsorbed water molecules on the substrate or the dissolved water in the organic solvent. Then, the hydroxylated OTS molecules are adhered to substrate. Finally, Si–O–M anchoring and lateral bonding are formed by the condensation between Si–OH and M–OH.

From the above example, it can be inferred that the water inducing the hydroxylation of surfactant is a key factor for the assembly of organosilane. So, it is clear that OTS molecules can only be physically absorbed on the substrate without the absorbed water or in the solvent without the dissolved water.

Temperature is another important factor affecting SAM formation (Ulman 1996). This can be attributed to the competition between the reaction of hydrolyzed (or partially hydrolyzed) trichlorosilyl (or alkoxysilyl) groups

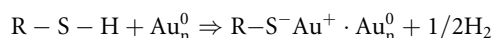


Solid-Like Lubricating Films, Self-Assembled Films, Fig. 1 A typical preparation process for SAMs (a); the molecular structure of a SAMs (b)

with other such groups in solution to form a polymer, and the reaction of such groups with surface M–OH moieties to form a SAM. As temperature increases, the surface reaction is suppressed and the thermal disorder in the forming monolayer increases. Thus, there must exist a critical temperature (T_c), below which an ordered monolayer can be formed. It is reported that the critical temperature of T_c is a function of chain length. Particularly, a linearity between T_c and the number of carbon (N_C) was observed when N_C is in the range of 10–22 (Ulman 1996).

Organosulfur Monolayers

Due to the excellent coordination to the transition metal surface, organosulfides, such as R–SH, R_1 –S– R_2 and R_1 –S–S– R_2 , can be strongly attached to the surface of Au/Ag/Cu/Pt/Hg/GaAs/InP to form ordered SAMs (Ulman 1996). However, the most studied SAMs are those of alkanethiols on Au (111) surfaces. In a typical assembling process, a fresh Au layer deposited on silicon or mica substrate was immersed into a solution (the solvent can be ethanol, hexane, acetone, or dichloromethane) of alkanethiol with a concentration of 10^{-3} – 10^{-1} mol/L for a period of time (several minutes to several days). In such cases, the interfacial reaction may be formally considered as an oxidative addition of the S–H bond to the Au surface, followed by a reductive elimination of the hydrogen. When a clean Au surface is used, the proton probably ends as an H_2 molecule (Ulman 1996):



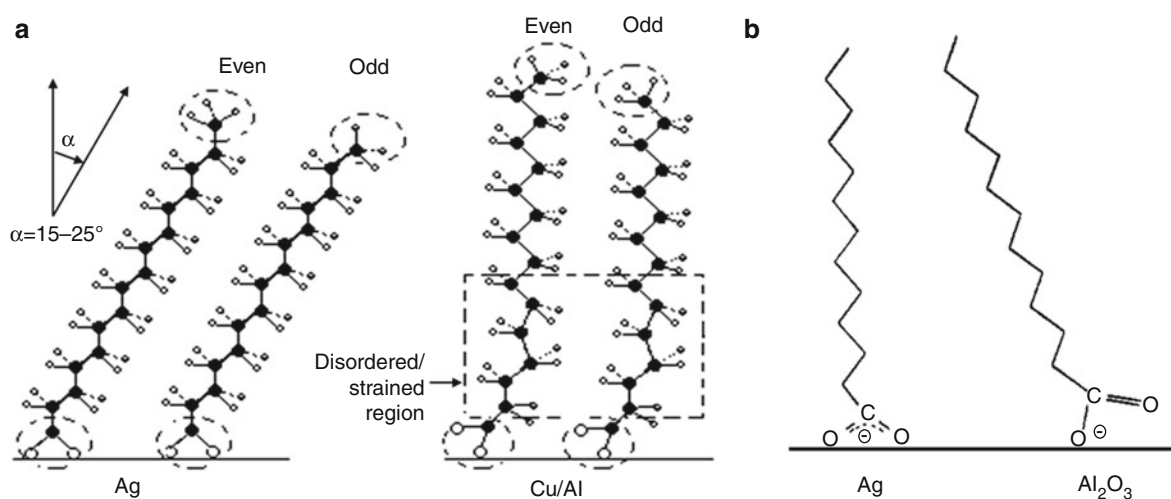
The adsorbing species of RS^- has been evidenced by characterizations of X-ray photoelectron spectroscopy (XPS), Fourier transform infrared (FTIR) spectroscopy, Raman spectroscopy, and electrochemistry. The

homolytic bond strength of such bonding is as high as ~ 40 kcal/mol.

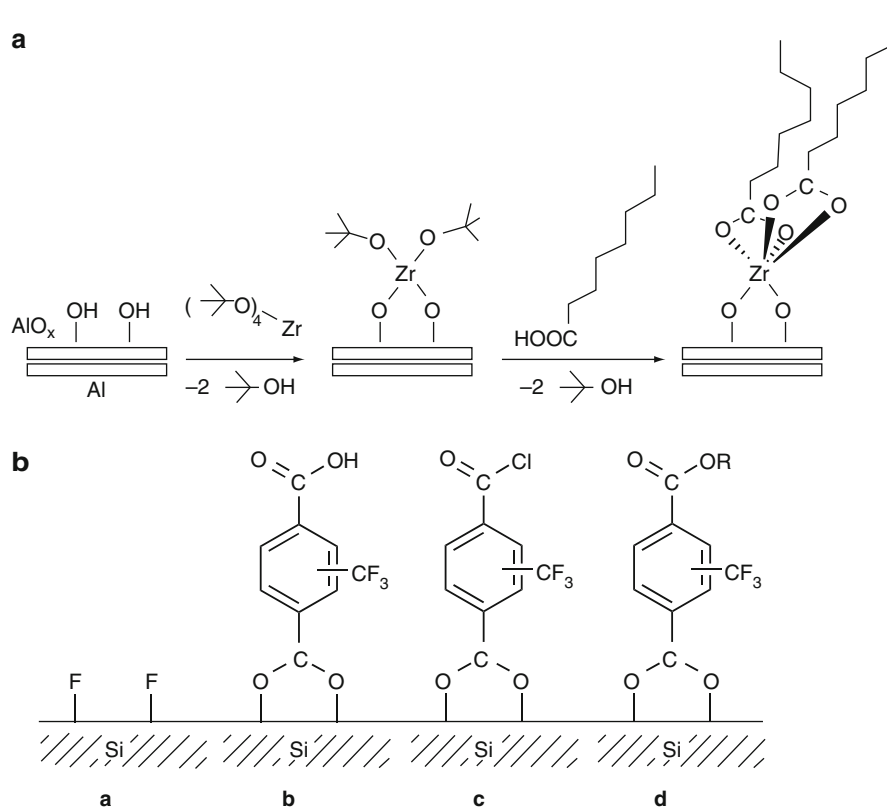
Fatty Acid Monolayers

Long-chain *n*-alkanoic acids ($C_nH_{2n+1}COOH$) can be spontaneously adsorbed onto the natively oxidized metal surfaces (such as Ag/Cu/Al) to form stable assemblies. The substrate of such metal oxide is positively charged and can be regarded as Lewis acid; the ionized *n*-alkanoic acid is a Lewis base. So, the self-assembly is an acid–base reaction. The driving force is the formation of surface salt between carboxylate anions and surface metal cations. The as-formed SAMs are very stable, with interface ionic bonding as high as 120 kJ/mol. It was discovered that the spatial structures of the as-formed interfacial ionic bonding are not unique. To be specific, as discovered by Tao (1993), the two oxygen atoms of carboxylate bind to the Ag surface almost symmetrically and the molecular chain extends trans zigzag, forming tile angles between 15° and 25° , while on the surface of Cu/Al the carboxylate binds asymmetrically to the surface and the molecular chains pack straight up for the long chain acid (Fig. 2a). Moreover, the orientation of tail groups (i.e., $-CH_3$) vary as the carbon number in alkyl chain changes from even to odd. However, as to the orientation of the alkyl chain, contrary phenomena have been observed by Pemberton et al. (Fig. 2b, Thompson and Pemberton 1995). Such a contradiction may be caused by the different smoothness of the substrates.

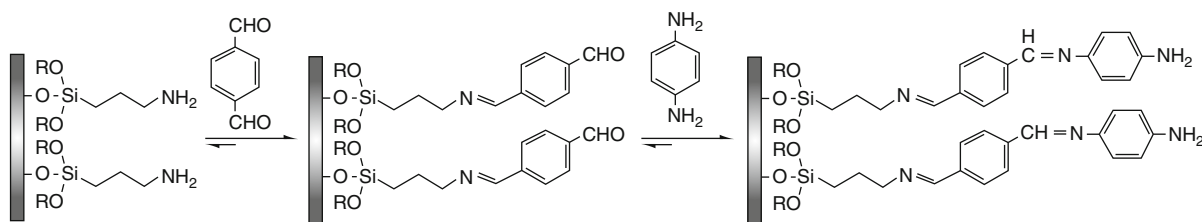
In some specific situations, interfacial covalent bonding also can be realized. For instance, mediated by an alkoxide of zirconium, *n*-octanoic acid can be covalently adsorbed onto the Al_2O_3 surface and the as-formed SAMs is stabilized (Fig. 3a, Aronoff et al. 1997). In the case of a fluorinated silicon surface, aromatic dicarboxylic acid



Solid-Like Lubricating Films, Self-Assembled Films, Fig. 2 The proposed structure of *n*-alkanoic acid monolayer on different substrate



Solid-Like Lubricating Films, Self-Assembled Films, Fig. 3 Covalent adsorption of *n*-octanoic acid onto the Zr alkoxide treated hydroxylated Al surface (a) and aromatic dicarboxylic acid onto the fluorinated silicon surface (b)



Solid-Like Lubricating Films, Self-Assembled Films, Fig. 4 Formation of π -conjugated azomethine oligomers by controlled iminization of aromatic dialdehydes and diamines onto an amino-functionalized substrate

can be covalently bonded and can serve as an active layer to induce further assembling (Fig. 3b, Mitsuya and Sugita 1997).

Self-Assembled Multilayer Films (SAMFs)

The exposed tail groups of SAMs not only determine the surface properties but also can serve as the reactive points to construct SAMFs. For example, the reactive tail groups, such as amine, carboxyl, hydroxyl, and epoxy, can undergo different reactions to fabricate various SAMFs. Taking advantage of the condensation between amine ($-\text{NH}_2$) and aldehyde ($-\text{CHO}$) groups (Fig. 4), SAMFs composed of aromatic dialdehydes and diamines can be formed on an amino-functionalized substrate (Dinglasan et al. 2002). Moreover, in Yang's group, a series of SAMFs has also been constructed based on the surface amidation between the amine tail group and carboxyl (COOH)/acid chloride (COCl) (Song et al. 2008; Ou et al. 2009).

The COOH tail groups of SAMs are generally modified by amino compounds through amidation. However, it is well known that the amidation between COOH and amine groups is difficult to realize under normal conditions. To boost such a reaction, a dehydrating agent such as 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide (EDC) is often added. Another effective way to improve the reactivity of COOH is to convert it into COCl . The COOH exposed to SOCl_2 steam can be converted into COCl , which then reacts with alkyl mercaptan to produce multilayer film (Kim et al. 1995).

The tail hydroxyl is another popular group that is often used to induce the interfacial reactions. As depicted in Fig. 5, the hydroxyl group can be further modified by different reagents to construct various multilayers (Ulman 1996).

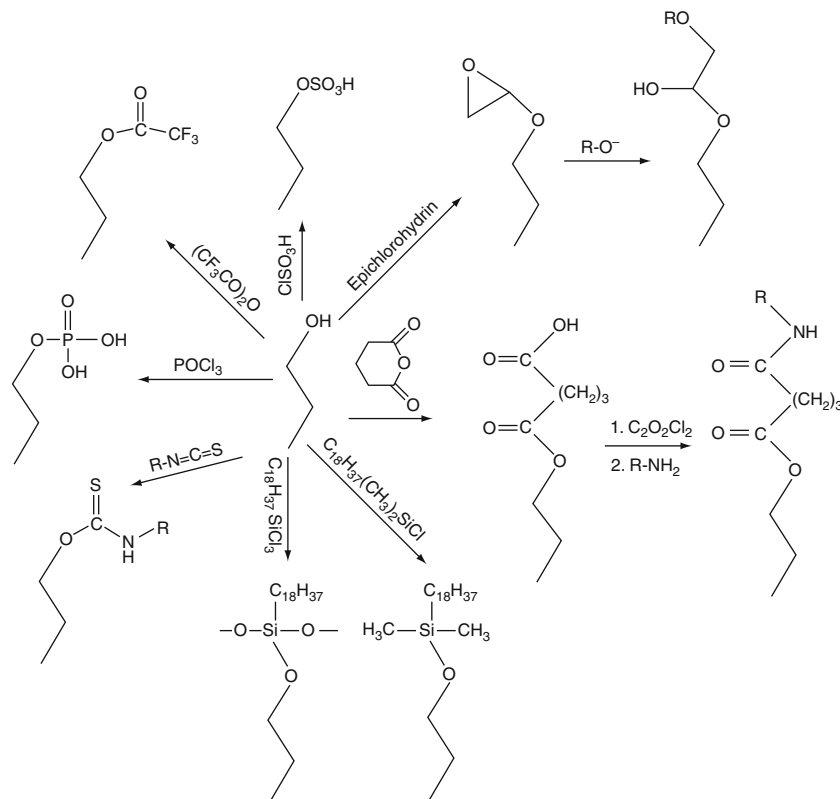
Gamma-glycidioxypropyl trimethoxysilane (GPTMS) SAM is widely studied as the adhesive layer to construct multilayer films based on the reactions between epoxy and amine/hydroxy/carboxy (Zhao et al. 2009). For example,

taking advantage of the ring-opening reaction between epoxy and amine groups, 3-aminopropyl triethoxysilane (APTES) was grafted onto the GPTMS-SAMs. Subsequently, OTS can be further grafted onto the APTES surface by the hydrolysis-condensation process. Thus, a trilayer film (coded as GAO) composed of GPTMS, APTES, and OTS was successfully obtained (Fig. 6).

To date, the most studied SAMFs are produced by an electrostatic layer-by-layer (ELbL) self-assembly technique, which is schematically shown in Fig. 7a. Briefly speaking, this ELbL involves immersion of a charged substrate into different solutions with oppositely charged materials. In addition to the electrolytes listed in Fig. 7b, other species (such as DNA, protein, nanoparticles of SiO_2 , TiO_2 , and ions of Tb^{3+} , Ce^{4+} , Sn^{4+} , and Zr^{4+}) have also been incorporated into the multilayered systems (Ariga et al. 2007).

However, this electrostatic LbL technique is confined to charged materials. To expand the application of the LbL, a novel non-electrostatic LbL (NELbL) assembly technique has been invented in Yang's group (Ou et al. 2011). The newly reported polydopamine (PDA) serves as the building block due to its special nature, i.e., high adhesion to almost all solid surfaces and the active surface with functional groups (such as $-\text{OH}$ and $-\text{NH}_2$). As schematically illustrated in Fig. 8, PDA can be chemically grafted onto the amine groups of APTES-SAMs (Fig. 8, Process II) or hydroxyl groups of ZrO_2 film (Fig. 8, Process IV). In addition, the ZrO_2 clusters formed in the $\text{Zr}(\text{SO}_4)_2$ solution can deposit onto the PDA surface via chelation (Fig. 8, Process III). Therefore, the sequential deposition of ZrO_2 and PDA can present a novel non-electrostatic strategy to construct ZrO_2 /PDA multilayer films.

On the other hand, the LbL procedure is generally performed in aqueous solution that is able to dissolve the charged materials. In 1997, Wang et al. (1997) developed a method to fabricate SAMFs in ethanol solution. It is found that the interlayer interaction is the hydrogen



Solid-Like Lubricating Films, Self-Assembled Films, Fig. 5 Construction of multilayers through surface reaction of hydroxyl-terminated monolayers with other groups

bonding between polyacrylic acid (PAA) and polyvinyl pyrrolidone (PVP). Thus, LbL can be expanded to the non-aqueous solution system.

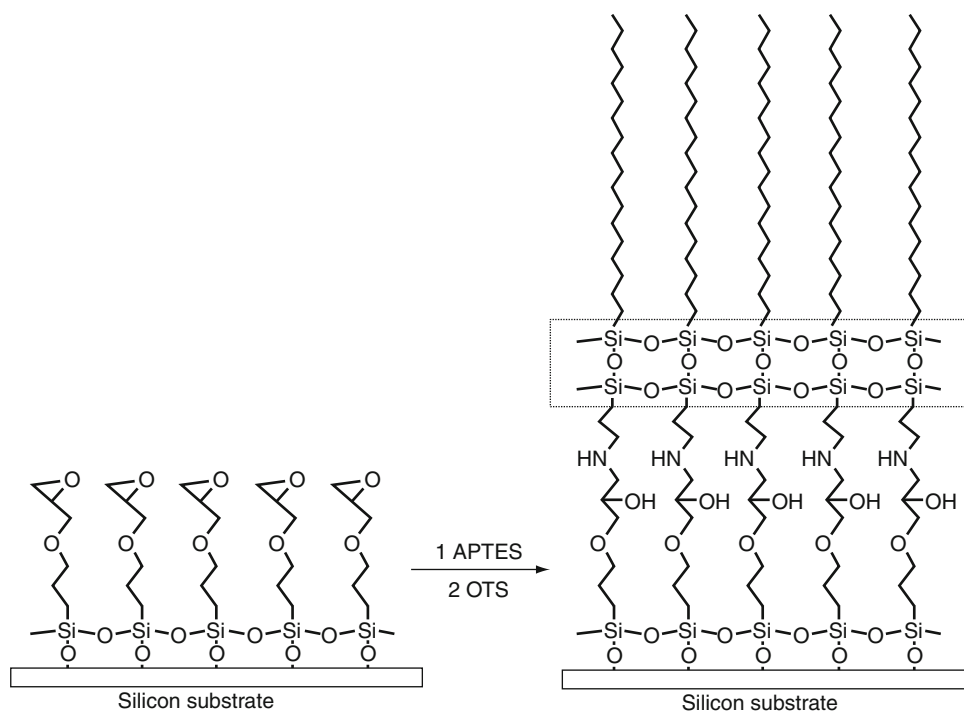
According to various interlayer interactions such as electrostatic force, hydrogen bonding, and covalent bonding, various SAMFs can be prepared successfully.

Key Application

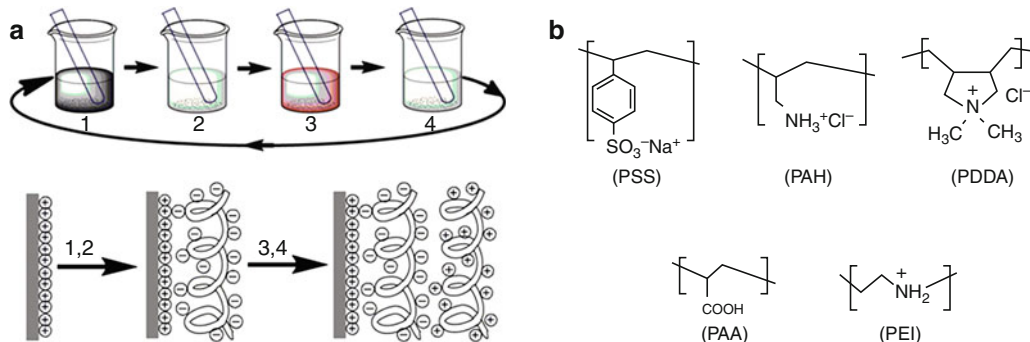
Lubricating film for micro-/nano-electromechanical systems (MEMS/NEMS)

Development of MEMS/NEMS has given rise to attempts to use ultrathin organic layers, such as SAFs, as lubricating coatings. MEMS/NEMS are featured by a diminishing gap between mating surfaces. For modern and future applications, this gap should be kept on a nanoscale level. However, owing to the miniaturization

of devices, serious stiction and friction problems occur. Treatment of surfaces by means of various coatings to reduce the stiction and friction is definitely desirable. SAFs are ideal candidates for lubricating film in MEMS/NEMS, not only for the dimension matching but also for its simple preparation process (Tsukruk 2001). Taking the well-studied OTS-SAMs as an example, SAFs suitable for lubricating film should meet some requirements; a strong interfacial binding (Si–O–M, M is the atom of the substrate) is needed to enhance the stability of the film, a long alkyl chain [–(CH₂)₁₇–] related to the load-carrying capacity is preferred. Moreover, the hydrophobic tail group (–CH₃) determining the adhesion/friction force is also necessary. It was found that OTS-SAMs possessed much lower friction coefficient and greater load-carrying capacity as compared with APTES-SAMs.



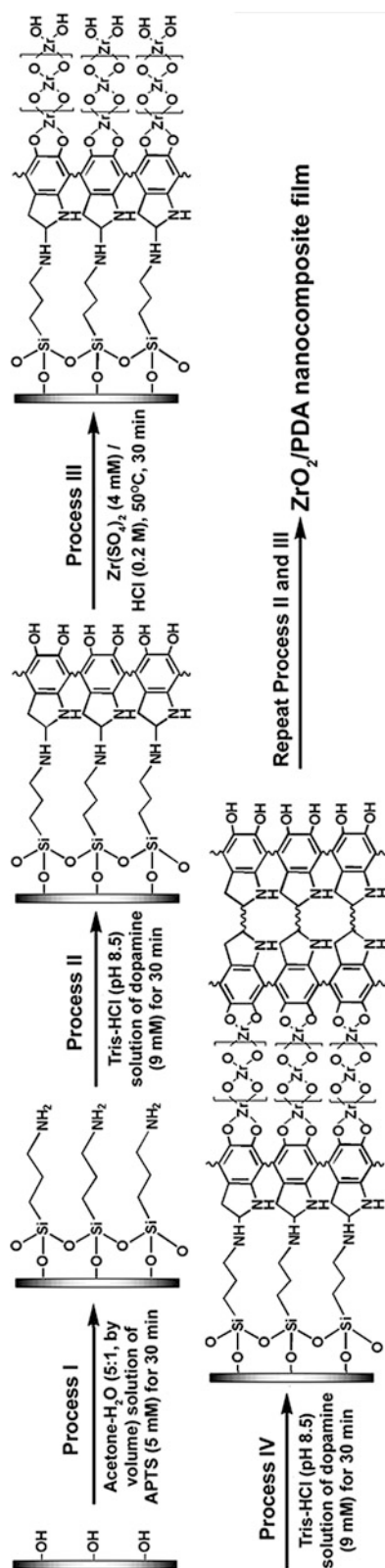
Solid-Like Lubricating Films, Self-Assembled Films, Fig. 6 Grafting of APTES onto the GPTMS-SAMs via epoxy ring opening and further modification with OTS to construct self-assembled multilayer film



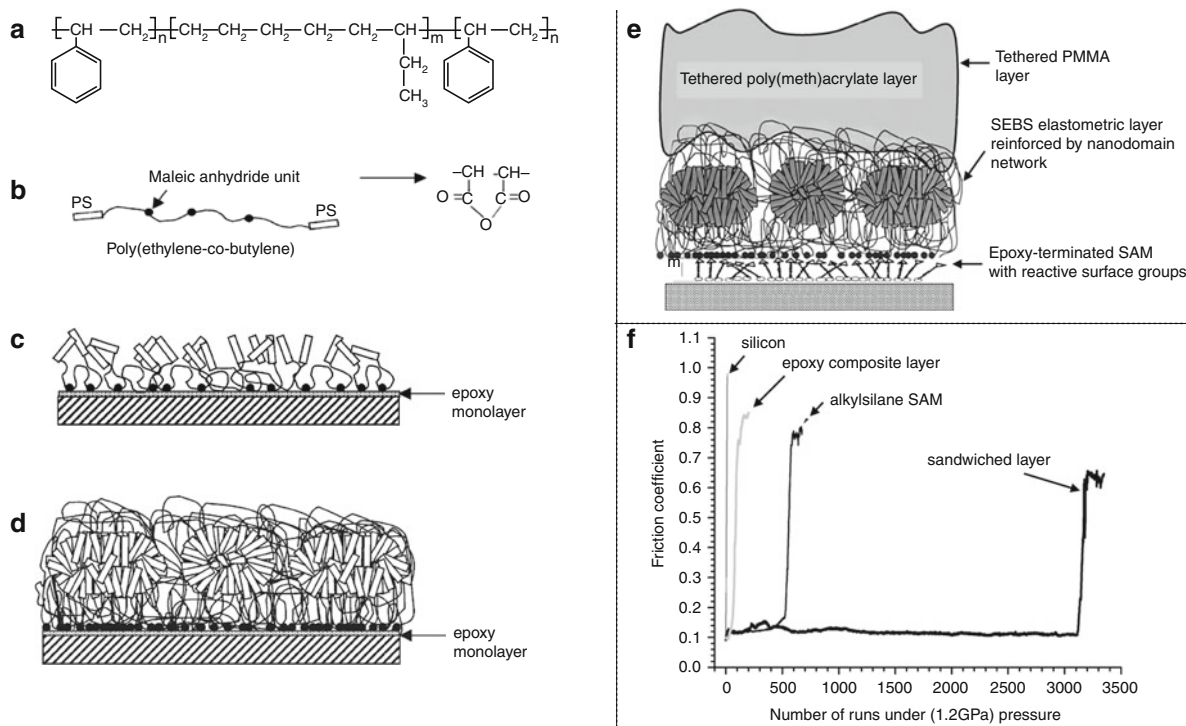
Solid-Like Lubricating Films, Self-Assembled Films, Fig. 7 The schematic view for the process of ELBL (a); chemical structures of the typical polyelectrolytes used in ELBL (b)

A lot of effort has gone into improving the tribological performance of SAFs by constructing various SAMFs. Especially, the interchain interaction within SAMFs is enhanced and the load-carrying capacity is expected to be improved

(Song et al. 2008). For example, a series of dual-layer SAMFs has been constructed and the interchain hydrogen bonding is thought to be responsible for the lengthened anti-wear life at a high applied load of 0.3 N (Song et al. 2008).



Solid-Like Lubricating Films, Self-Assembled Films, Fig. 8 A schematic view for constructing ZrO_2/PDA SAMF



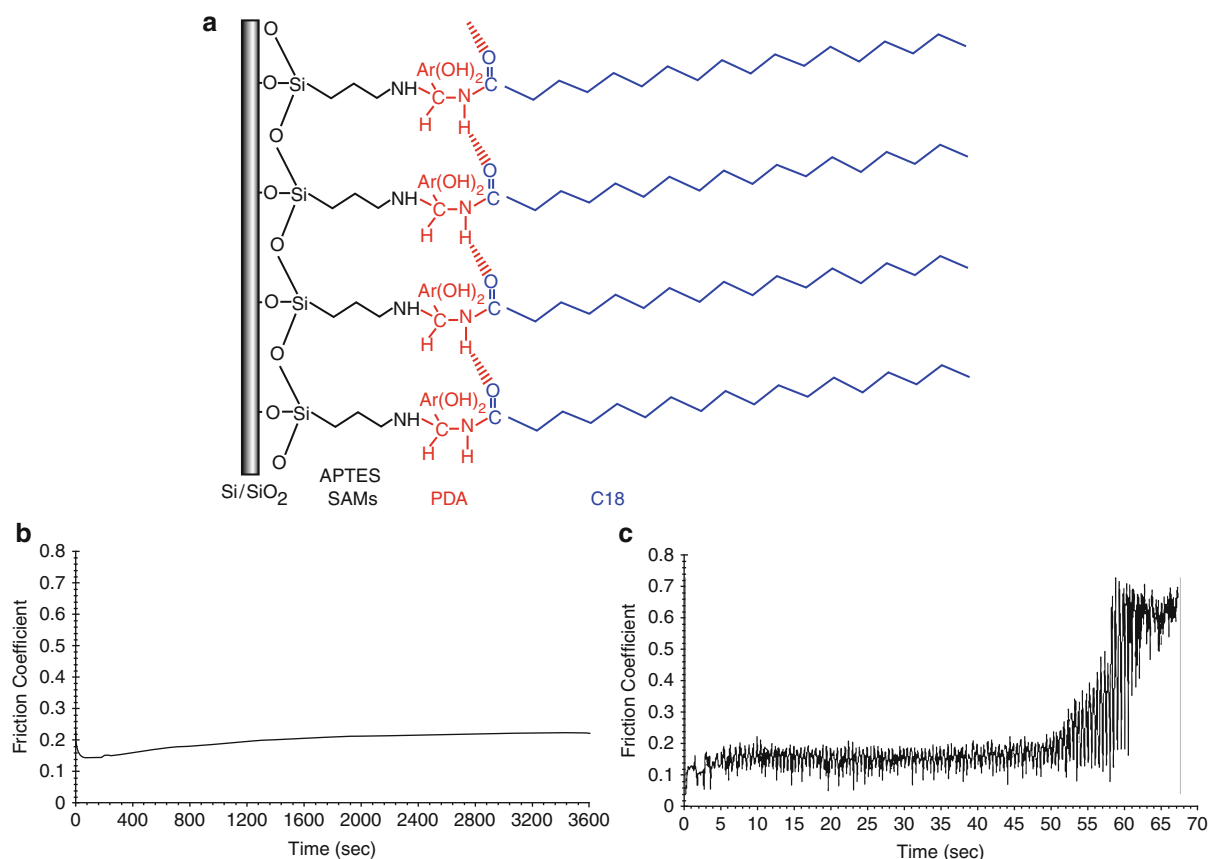
Solid-Like Lubricating Films, Self-Assembled Films, Fig. 9 Chemical (a) and schematic (b) structure of SEBS; SEBS layer with disordered structures and a thickness <2.5 nm (c); SEBS layer with nanodomain morphology and a thickness <2.5 nm (d); Architecture of sandwiched trilayer (e); Friction coefficient versus the number of reciprocal sliding runs for different samples (f)

The polymer-based SAFs with cross-linking network structures can sustain high compression and shear stress. For example, a copolymer of poly[styrene-*b*-(ethylene-co-butylene)-*b*-styrene] (coded as SEBS, Fig. 9a, b) functionalized with 2% maleic anhydride into the hydrocarbon chains was assembled onto the surface of epoxy-terminated monolayer (Fig. 9c, d, Luzinov et al. 2001). The as-fabricated films possess low friction coefficient, modest adhesion, low stiction, and good wear stability as well. To further improve the wear resistance, a SAF with trilayer sandwiched architecture has been constructed (Fig. 9e, Sidorenko et al. 2002). As expected, the anti-wear life is much longer than that of the epoxy composite layer (Fig. 9f).

Most recently, inspired by the structures of SAMs, a tri-layer film composed of an APTES underlayer (behaving as the head group in SAMs), a PDA interlayer (as the alkyl chain in SAMs), and stearoyl chloride (as the tail

group in SAMs) was prepared. The carrying-capacity of this tri-layer film was further boosted as compared with the corresponding dual-layer SAMFs without PDA interlayer (Fig. 10, Ou et al. 2009).

As discussed above, various SAFs with properties of low friction and good wear resistance have been investigated extensively as lubricating coatings for MEMS/NEMS. These fundamental works have generated the basic understanding of microtribology. To develop practical lubricating SAFs in MEMS/NEMS, however, there are still challenges to be overcome. For example, to date, most of the lubricating SAFs studied have been constructed on a smooth silicon substrate, whereas the actual surface in MEMS/NEMS is much more complicated and multifarious. Therefore, more versatile assembling processes and systems should be developed and more work is needed.



Solid-Like Lubricating Films, Self-Assembled Films, Fig. 10 A schematic view for the formation and combination bonding of the three-layer film on silicon wafer (a); The macrotribological behaviors of the tri-layer SAMFs (b) and the dual-layer SAMFs without PDA interlayer (c)

Cross-References

- ▶ [Anti-Adhesion/Stiction Surface Design, Fabrication, and Applications](#)
- ▶ [Asperities](#)
- ▶ [Atomic-Level Stick-Slip](#)
- ▶ [Bonding at Surfaces/Interfaces](#)
- ▶ [Capillary Force and Surface Wettability](#)
- ▶ [Polymer Nanolayers](#)
- ▶ [Self-Assembled Monolayers](#)
- ▶ [Surface Force Apparatus](#)
- ▶ [Surface Forces, Surface Tension, and Adhesion](#)
- ▶ [Surface Free Energy](#)
- ▶ [Surface Roughness](#)
- ▶ [X-Ray Photoelectron Spectroscopy \(XPS\)](#)

References

- K. Ariga et al., Layer-by-layer assembly as a versatile bottom-up nanofabrication technique for exploratory research and realistic application. *Phys. Chem. Chem. Phys.* **9**, 2319–2340 (2007)
- Y.G. Aronoff et al., Stabilization of self-assembled monolayers of carboxylic acids on native oxides of metals. *J. Am. Chem. Soc.* **119**, 259–262 (1997)
- J.A. Dinglasan et al., Formation of conjugated azomethine oligomers on quartz and silicon oxide surfaces. *Langmuir* **18**, 5971–5973 (2002)
- T. Kim et al., Polymeric self-assembled monolayers. 2. Synthesis and characterization of self-assembled polydiacetylene mono- and multilayers. *J. Am. Chem. Soc.* **117**, 3963–3967 (1995)
- I. Luzinov et al., Nanotribological behavior of tethered reinforced polymer nanolayer coatings. *Tribol. Int.* **34**, 327–333 (2001)
- M. Mitsuya, N. Sugita, Chemisorption of dicarboxylic acids on an Si(111) surface and subsequent chemical reactions at the surface of adsorbed molecular layers. *Langmuir* **13**, 7075–7079 (1997)
- J.F. Ou et al., Self-assembly and tribological property of a novel 3-layer organic film on silicon wafer with polydopamine coating as the interlayer. *J. Phys. Chem. C* **113**, 20429–20434 (2009)
- J.F. Ou et al., Mechanical property and corrosion resistance of zirconia/polydopamine nanocomposite multilayer films fabricated via a novel non-electrostatic layer-by-layer assembly technique. *Surf. Interface Anal.* **43**, 803–808 (2011)
- A. Sidorenko et al., Wear stability of polymer nanocomposite coatings with trilayer architecture. *Wear* **252**, 946–955 (2002)

- S.Y. Song et al., Preparation and tribological behaviors of an amide-containing stratified self-assembled monolayers on silicon surface. *Langmuir* **24**, 105–109 (2008)
- Y.T. Tao, Structural comparison of self-assembled monolayers of N-alkanoic acids on the surfaces of silver, copper, and aluminum. *J. Am. Chem. Soc.* **115**, 4350–4358 (1993)
- W.R. Thompson, J.E. Pemberton, Characterization of octadecylsilane and stearic acid layers on Al_2O_3 surfaces by Raman spectroscopy. *Langmuir* **11**, 1720–1725 (1995)
- V.V. Tsukruk, Molecular lubricant and glues for micro- and nanodevices. *Adv. Mater.* **13**, 95–107 (2001)
- A. Ulman, Formation and structure of self-assembled monolayers. *Chem. Rev.* **96**, 1533–1554 (1996)
- L.Y. Wang et al., A new approach for the fabrication of an alternating multilayer film of poly(4-vinylpyridine) and poly(acrylic acid) based on hydrogen bonding. *Macromol. Rapid. Commun.* **18**, 509–514 (1997)
- J. Zhao et al., Preparation and tribological studies of self-assembled triple-layer films. *Thin Solid Films* **517**, 3752–3759 (2009)

Solid-Liquid Bi-phase Lubricating Coatings

M. KALIN

Faculty of Mechanical Engineering,
University of Ljubljana, Ljubljana, Slovenia

Synonyms

Adhesion and lubricious coatings; Adsorption; Boundary lubrication by lubricious coatings; Friction and coatings; Lubrication through coatings; Lubricious diamond-like carbon coatings; Wear of coatings

Definition

In spite of the low chemical activity and “inertness” of DLC coatings, it has become clear in recent years that some DLC coatings react with additives and form various tribochemical products and films. However, these interactions are weaker, the tribofilms are thinner, and they are less wear resistant in comparison to steels. At the same time, the importance of the various physical effects of the fluid (oil) and the well-known, wear-protective, and low-friction nature of the DLC coatings themselves with regard to tribological behavior are almost totally unclear. In other words, the roles of the different “components” in the boundary lubrication of DLC coatings, i.e., the contribution of the additives, the base oils, and the DLC coating itself are not fully understood. Therefore, in this work an attempt to elucidate some of the aspects of these questions

related to the physical properties of base oils, the adsorption of oil films and their strength, the DLC coatings and avoiding the chemical effects of the oils or additives is presented. Results show a dramatic improvement with the use of a base oil – particularly when it has high viscosity – on DLC coatings’ wear and durability, but just the contrary – a friction increase – when compared with non-lubricated DLC contacts. Based on some specific quasi-static experiments it is also suggested that base oil films might adsorb at the DLC surface and protect it against wear.

Scientific Fundamentals

Introduction

Modern mechanical systems need to operate under demanding working conditions, such as high loads, high speeds, high temperatures, and adverse environments. Furthermore, in milder conditions better results are expected today than in the past. Sometimes reduced production costs are also required, for example, by reducing the amount of surface finish. Under lubricated conditions, which represent the vast majority of all mechanical systems, this suggests that lubrication is changing from the hydrodynamic to the mixed or boundary regime, where direct contacts between surfaces are dominant. In addition, many systems regularly operate under conditions of poor lubrication, due to starved lubrication or the nature of the loading conditions. The successful operation of tribological contacts and thus the whole system is therefore dependent on the properties of the surfaces and the ability to form wear-protective and low-shear, tribochemical, boundary-interface films.

Around the world there are various attempts to solve these problems, including the development of better oil additives, new wear-reducing and low-friction coatings, multi-layered and “smart” nano-structured coatings, chameleon surfaces, and textured surfaces. The boundary lubrication of DLC coatings is among the efforts toward addressing these problems (Matthews et al. 1998; Erdemir and Donnet 2006; Velkavrh et al. 2008; Kalin et al. 2008; Neville et al. 2007). DLC coatings are known for their high hardness, low-friction properties, good wear resistance, and, typically, poor wetting and “inertness” (i.e., a low surface energy) (Sanchez-Lopez et al. 2003; Kalin et al. 2009). Inertness, which is an important property for low oxidation, corrosion, and adhesion prevention, is also a drawback for successful boundary lubrication, where interactions and tribochemical reactions between the surface and the lubricants are required to ensure low friction and low wear. Namely, oil additives that are currently

available are designed for particular metal engineering surfaces, which are polar and/or can form reactive nascent surfaces under severe conditions, so that the reactions are possible (Mortier and Orszulik 1993). In contrast, pure DLC coatings (a-C:H, a-C) have few active sites, thus they cannot readily react with additives and oil molecules. Accordingly, the efficiency of the actual boundary “lubrication” with DLC coatings is still under investigation and many times DLC coatings are only considered as a passive member in the contact (usually with steel present), and the steel counter-body, i.e., steel, takes the “lubrication” part in the contact – at least according to conventional theories.

There is little understanding of the details of the lubrication mechanism of DLC coatings. This area is relatively new and researchers have used very different coatings with different physico-chemical properties, hardness, roughness, etc., as well as using a variety of different oils and additives at different concentrations, with different chemistries. Tests have been performed at different loads, velocities, and temperatures. Thus, results are very often difficult to compare and sometimes they are even contradictory.

So far, it is mainly the chemical effects of additives on DLC coatings that have been investigated. In spite of the above-mentioned difficulties, clear evidence for chemical reactions between DLC coatings and oil additives has been presented (De Barros’Bouchet et al. 2005; Kalin et al. 2007, 2010; Equey et al. 2007). It appears that some DLC coatings react with additives and form various tribochemical products and films. Indeed, these interactions are weaker, and the tribofilms are thinner and less wear resistant, and, thus, the reactivity of DLC coatings (with the additives present) is certainly lower compared with steels. Some possible mechanisms for the films’ formation were also presented in these studies; however, generally valid concepts and boundary-lubrication mechanisms, similar to those known for metals, are still missing.

At the same time, it is almost totally unclear as to what is the role of the oil – as a base fluid – in the DLC contacts and how important, in terms of the tribological behavior, are the chemical interactions with the base oils and additives in comparison with the physical effect of the fluid (oil) and the wear-protective and low-friction nature of the DLC coatings themselves. Thus, what is the actual overall role of the different “components” in oil boundary lubrication, i.e., the contribution of the additives, the base oils, and the DLC coating itself?

In the literature, there is a severe lack or even an absence of studies that focus on the physical effects of the oils (Kalin et al. 2009; Velkavrh et al. 2009) or the interactions between the base oils and the DLC coatings, i.e., boundary lubrication based on adsorption

mechanisms (Ozbolt et al. 2009). However, adsorption is a very important lubrication mechanism at lower temperatures or in cases where more chemically active and corrosive additives are not desired, not acting, or not available. Furthermore, lubrication based on adsorption is also more sensitive to contact severity, and when the adsorption films break down in contacts, the surface properties will define the performance of these contacts. Therefore, there is an important interplay between the surface property and the strength of the adsorption on the overall performance of such contacts. This seems to be especially important and promising in the case of DLC coatings, where these coatings also show excellent performance under dry conditions (Erdemir and Donnet 2006; Erdemir 2004).

It is important to answer the following questions:

- What is the role of the base oil, as a fluid, in the boundary-lubricated contacts?
- Do base oils form adsorption layers at DLC coatings?
- What is the effect of these layers – if they appear – on the tribological performance of boundary-lubricated DLC coatings?
- What are the individual contributions of the DLC coatings themselves and what is the contribution of the base oil to the tribological behavior of boundary-lubricated DLC/DLC contacts?

In this work it is intended to elucidate some aspects of the above questions. Only base oils have been used, so as to exclude any additive chemical effects and focus only on the base oils’ influences, which are mainly physical ones. Since tribological in-situ adsorption effects, especially physical-based effects, are difficult to prove and analyze with ex-situ techniques, the concept of comparing the behavior of DLC coatings with the well-known, boundary-lubrication effects of steels under the same conditions using variations based on a Stribeck parameter was employed. The tests were performed at very low velocities so as to correspond to boundary lubrication. In addition, specific quasi-static tests with even lower velocities ($\mu\text{m/s}$ range) to allow further investigations of the adsorption of the base oils at the DLC surfaces were also performed. Moreover, to understand the effect of the base oil (as a fluid) in boundary-lubricated DLC contacts – compared with non-lubricated contacts under the same conditions – experiments under non-lubricated conditions have also been performed.

Experimental Details

Tribological Tests and Surface Analyses

The wear experiments, in which only self-mated, steel/steel, and DLC/DLC contacts were used in order to

exclude the influence of the combined (and thus undefined) effects of different counter-body materials, were performed in a reciprocating sliding device. The disks were fixed in the base, while the upper specimens, i.e., the balls, were fixed in the oscillating holder. In all the experiments, 10 N of normal load was applied through the loading system, which resulted in an initial Hertzian contact stress of about 700 MPa (1 GPa max.). A stroke of 13.6 mm and various oscillating frequencies were set to ensure various relative contact velocities from 0.003 to 0.04 m/s. In each test the total sliding distance was set to 100 m, corresponding to 7,350 loading cycles. These tests were thus varied by velocity at a fixed load and viscosity, to satisfy the evaluation according to Stribeck-curve behavior. Furthermore, some experiments were performed at an even lower velocity, i.e., at 500 $\mu\text{m/s}$, referred to in this work as quasi-static because of the extremely low velocity. In most experiments the oil was applied to the surface prior to the tests, although in some cases the tests were performed without any lubrication at room temperature and in room-humidity conditions. On the basis of these contact conditions, in lubricated tests the calculated Lambda value was always below 0.2, even for the highest velocity used, which suggests that the conditions used were always within the boundary-lubrication regime. This was also true after the running-in and the smoothing of the surfaces.

Frictional force was monitored throughout the test and was digitally recorded. Before the test the specimens were ultrasonically cleaned in high-purity benzene and ethanol, and a small amount of oil was spread on the surface of the flat specimen prior to each experiment. After the test, the specimens were carefully cleaned in ethanol, to remove any residual oil, and then dried in a stream of air. The wear volume was calculated using the ball's wear-scar diameter, measured with an optical microscope (Leitz Miniload 2, Ernst Leitz Wetzlar, GmbH, 6330 Wetzlar, Germany) and using a geometrical equation for the volume of a spherical segment. Measurements were made on each wear scar in the parallel and perpendicular directions, and the mean values of these measurements were used in the wear-volume calculations.

After the experiments, the worn surfaces were analyzed by employing stylus-tip profilometry (T8000, Hommelwerke GmbH, Schwenningen, Germany), and scanning electron microscopy (SEM; JEOL JSM-T330A) combined with a light-element Si energy-dispersive (EDS) detector (beryllium-window type) at an accelerating voltage of 20 kV and an Inca Energy data-processing unit (Oxford Instruments Analytical Ltd., UK). In addition, a number of samples tested under various conditions

were analyzed using an atomic force microscope (AFM) Veeco CP-II, using the contact mode with constant normal load and a scan size of $50 \times 50 \mu\text{m}$.

Substrate Material, Coatings, and Lubricants

The balls and the discs were made from AISI 52100/DIN 100Cr6 steel. All the balls and discs initially had the same mechanical, thermal, and surface characteristics. The steel balls were standard balls with a diameter of 10 mm and the steel flat samples were $\phi 24 \times 7.9$ mm discs, both having a hardness of 850 HV (corresponding to 8.3 GPa), measured with a microhardness tester (Leitz Miniload, Wild Leitz GmbH, D-6330 Wetzlar, Germany). All the steel balls and discs were coated with a single-layer of "pure," hydrogen-containing, amorphous, diamond-like carbon (a-C:H), having a thickness of $1.78 (+/-0.09) \mu\text{m}$. A RF-PACVD process at a frequency of 13.56 MHz was employed for the deposition of the coating, and a Si-based interlayer was used to improve the adhesion properties of the coating. The adhesion of the coatings was investigated with a scratch tester (Revetest, CSM Instruments SA, Switzerland), and the average Lc1 (initiation of forward chevron cracks at the borders of the cracks), Lc2 (initiation of forward chevron cracks at the borders of the cracks accompanied by interfacial spallation), and Lc3 (initiation of gross interfacial shell-shaped spallation) values were determined as 40.1 N, 56.7 N, and 71.6 N. The hardness and the Young's modulus of the coating were measured using the depth-sensing indentation technique (NanoTest 600 instrument with Berkovich indenter, Micro Materials Limited, UK). The hardness of the coating was 21.9 GPa, and the Young's modulus was 15.7 GPa. The final surface roughnesses, R_a and R_q , of the samples after the coating deposition, measured using a stylus-tip profilometer (T8000, Hommelwerke GmbH, Schwenningen, Germany), were about 0.03 and 0.05 μm for the balls and 0.05 and 0.07 μm for the discs, respectively.

Three different lubricants were used in the experiments: (1) a polyalphaolefin base oil of viscosity grade ISO VG 18 (denoted as PAO4), (2) a polyalphaolefin base oil of viscosity grade ISO VG 30 (denoted as PAO6), and (3) a polyalphaolefin base oil of viscosity grade ISO VG 46 (denoted as PAO8). Some of the key properties of the base oils are presented in Table 1.

Results

Non-lubricated Conditions

The experiments under non-lubricated conditions showed significant differences in the friction behavior between the steel and the DLC contacts (Fig. 1).

Solid-Liquid Bi-phase Lubricating Coatings, Table 1 Kinematic viscosity at 40 in 100°C and density at 20°C of polialfaolefin oils

Base oil	Kinematic viscosity @ 40°C (mm ² /s)	Kinematic viscosity @ 100°C (mm ² /s)	Density @ 20°C (kg/m ³)
PAO4	18	4	820
PAO6	30	6	828
PAO8	46	8	830

The same velocities were selected in these tests so as to be comparable with the velocities corresponding to boundary lubrication in the lubricated tests that are shown later.

It can be seen that in the steel/steel contacts the coefficient of friction was very high, between 0.65 and 0.75. This is a well-known and expected result due to the high contact pressure of 1 GPa (at the asperities it is even higher), which results in strong metallic bonds and adhesion between the two steel surfaces and, consequently, in high friction. An increase in friction of about 10% was measured as the velocity increased from 0.003 to 0.04 m/s (Fig. 1a).

The coefficient of friction for the non-lubricated DLC/DLC contacts is presented in Fig. 1b. The scale is the same as for the steel/steel experiments in order to show clearly the significant reduction of friction with dry DLC/DLC sliding. It is thus very obvious that the friction was significantly lower compared with the steel contacts, i.e., only about 0.04, which is 15–20 times less. However, this friction is only reported for the testing period when the coating was still covering the surface. Namely, for all the tested, non-lubricated conditions the DLC coating was worn through or spalled and then the friction increased due to the metallic contacts from the interlayer and/or the steel substrate. Moreover, the removal of the coating occurred more quickly at higher velocities, as is evident from Fig. 2b.

Figure 2 presents the wear results from the non-lubricated tests. For the steel/steel contacts the wear was measured after 100 m of sliding, corresponding to 7,350 cycles. A clear wear increase of about 50% with the velocity increase in the selected range can be seen in Fig. 2a. In contrast, the wear of the DLC/DLC contacts is not presented in terms of the volume loss because all the coatings wore through, and thus the measurements would be unrealistic. Instead, the number of cycles required to wear through the coatings is plotted. However, the same phenomenon can be seen as with the steel

contacts. Namely, the wear increased with the velocity and so the wear through or spalling occurred more quickly (Fig. 2b). This can be explained by the higher impact energy at the collisions of the asperities. The effect of the increased asperity-impact energy is even more pronounced with the DLC coating than with the steel (in spite of the 50% wear increase with the velocity increase) due to the lower fracture toughness and the reduced ability to plastically deform, which results in the cracking and partial spallation of the coating (starting at direct asperity contacts).

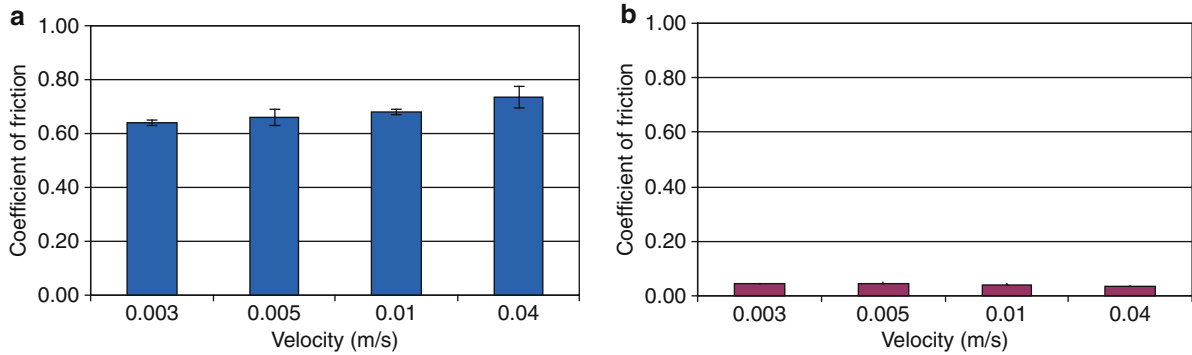
The above-presented steel-wear results are to be expected because it is known that dry steel contacts always lead to high adhesion and asperity deformation if the load is as high as 1 GPa. The steel surfaces, after the removal of thin oxide layers at high pressure, are prone to interact with each other and form strong adhesive bonds and material transfer. The velocity increase only stimulates these phenomena, so the wear increased by 50% in the selected velocity range (Fig. 2a).

In contrast, the wear of DLC coatings could be very different, depending strongly on the different contact conditions, as summarized in some reviews (Erdemir and Donnet 2006; Velkavrh et al. 2008; Kalin et al. 2008; Neville et al. 2007). Sometimes the wear was reported as very low, in particular when the sliding is unidirectional and the surfaces are very smooth. However, if the asperity contacts experience high contact stresses, such as in the case of rough surfaces or when most of the load is carried over a small number of asperities (dry contacts under high load), then the micro-fracture process initiates the initial surface damage and the wear or coating fracture progresses much more quickly. The increase in the velocity only accelerates this process and the failure occurs faster. This was very obviously the case of the present experiments. At a velocity of 0.003 m/s the coating failed after 6,600 cycles, while at 0.04 m/s it failed in less than 500 cycles (Fig. 2b). In agreement with this, the DLC surface damage with clear spallation, both on the ball and the disc, can be seen in Fig. 3.

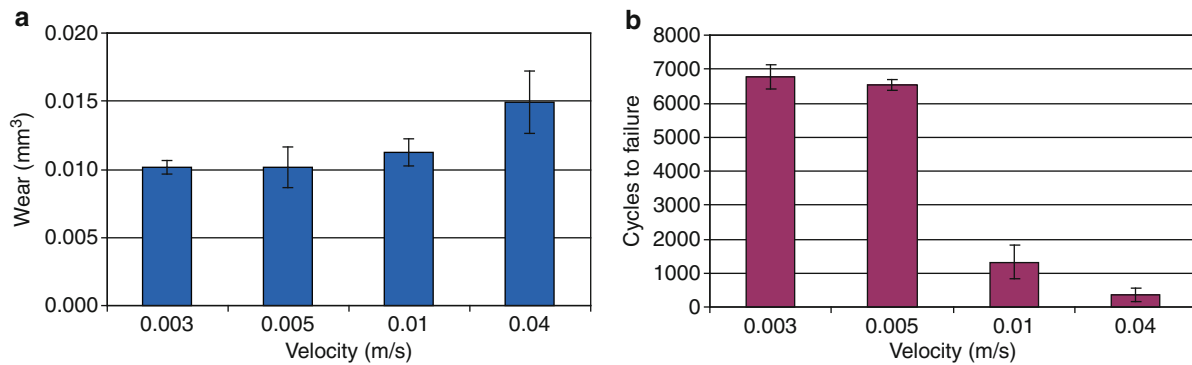
Lubricated Conditions

When the experiments were performed at the same velocities as in the dry conditions, i.e., from 0.003 up to 0.04 m/s, which corresponds to boundary lubrication, the friction and wear behavior changed dramatically in both cases, i.e., for the steel/steel and the DLC/DLC contacts.

In the steel contacts the friction reduced from about 0.7, measured in dry contacts, to about 0.14 in the boundary-lubrication regime, which is about a fivefold reduction. In addition, the friction became much more



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 1 Coefficient of friction for (a) steel/steel and (b) DLC/DLC contacts under non-lubricated conditions as a function of velocity



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 2 (a) Wear of ball in steel/steel contacts and (b) number of cycles to failure of DLC/DLC contacts under non-lubricated conditions as a function of velocity

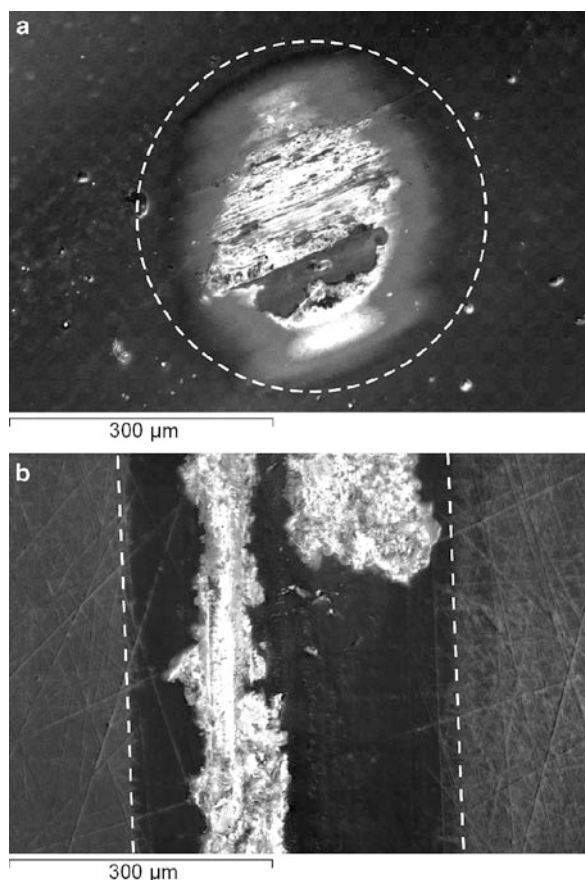
stable, and the curve was smooth and was almost at a constant value (Fig. 4a).

In contrast, the friction in the DLC/DLC contacts increased from about 0.04–0.11 (of course, the friction in the dry conditions was considered when the coating was still covering the surface). Figure 4b shows the lubricated and non-lubricated friction traces during the 100 m of sliding. The “dry” curve shows first a very low and stable friction, but then a dramatic increase and subsequently a very unstable friction of between 0.2 and 0.4 was observed during our experiments. In contrast, the lubricated friction curve is very smooth and stable, although at a somewhat higher value than the initial dry sliding.

Therefore, the increased friction due to the use of base oil in the DLC contacts may seem to make the tribological conditions in DLC/DLC contacts worse; however, a completely different conclusion must be drawn from the wear data (Fig. 5). Namely, the wear of the DLC/DLC in the lubricated tests became almost negligible

and the spallation (catastrophic wear) of the coatings never appeared, which was, however, always the case in the non-lubricated DLC/DLC contacts. The lubricated DLC/DLC wear was also about five times lower compared with the wear of the lubricated steel/steel contacts (Fig. 5). A similar conclusion can be drawn from the steel/steel contacts, where the wear was reduced by about 20 times (Fig. 5). These results also explain the dry friction curve of the DLC/DLC contacts in Fig. 4b. Namely, the friction first increased abruptly after the coating was removed, and then the continuation of the unstable friction depends on the actual contact situation (the extent of the metal and coating contacts).

Therefore, the base oil has a similar, strong, wear-protective effect on the DLC coatings as on the steel, but the reduction was quantitatively and qualitatively different. Namely, it is obvious that the extent of the wear cannot be directly compared due to different wear mechanisms, being mostly fracture-dominated (or at least

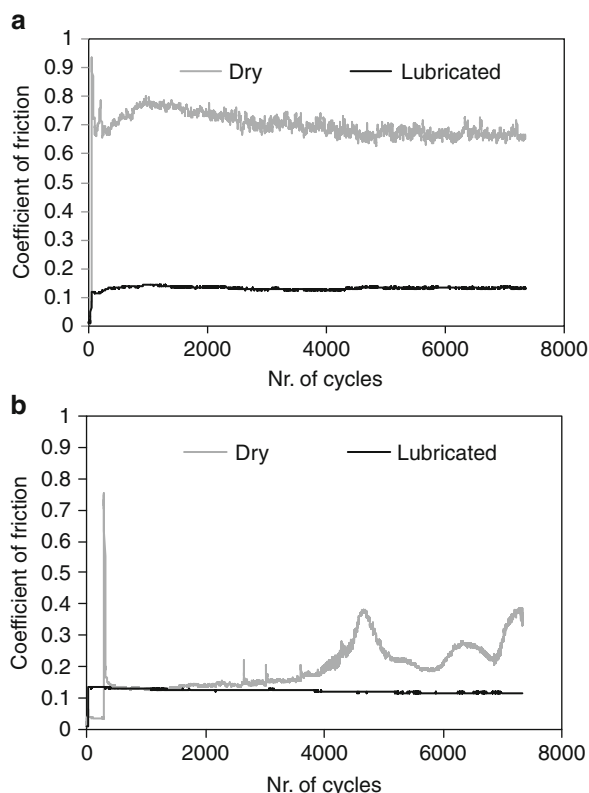


Solid-Liquid Bi-phase Lubricating Coatings, Fig. 3
Coefficient of friction for (a) steel/steel and (b) DLC/DLC contacts under non-lubricated conditions as a function of velocity

initiated) for the DLC and predominantly plastic in the case of the steel. On the other hand, lubrication drastically decreased the friction in the steel contacts (due to adhesion), but – in contrast – increased it in the DLC contacts (the absence of adhesion). However, the results clearly indicate the beneficial effect of the base oil on the durability of the DLC surfaces. Moreover, it should be noted that both wear and friction were lower in the lubricated DLC contacts compared with the lubricated steel contacts (Figs. 4 and 5).

Effect of the Base Oil's Viscosity

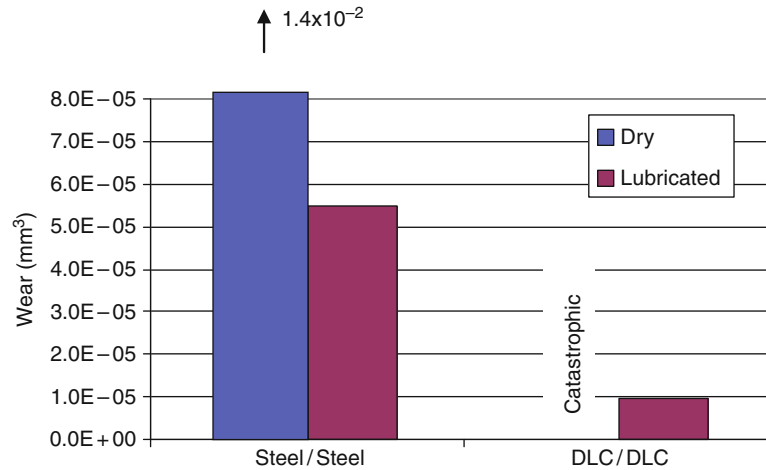
Viscosity is an important parameter that is directly proportional to the oil film's thickness and strength, and thus indicates its protective nature in tribological contacts. From lubricated tests we can conclude that the oil significantly reduced the wear (Fig. 5), and thus higher-viscosity



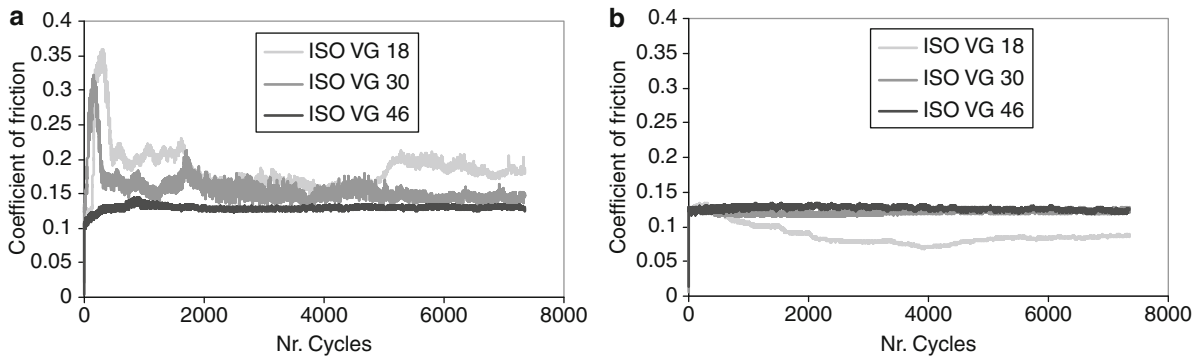
Solid-Liquid Bi-phase Lubricating Coatings, Fig. 4
Evolution of coefficient of friction for (a) steel/steel and (b) DLC/DLC contacts under non-lubricated and lubricated (using PAO4) conditions as a function of the number of cycles at a sliding velocity of 0.04 m/s

oil should provide an even more pronounced effect and better wear protection. The effect on friction is, however, more complex, but at such low velocities as used in our tests, certainly no hydrodynamic effects could be claimed.

Figure 6a shows the coefficient-of-friction curves for the steel/steel contacts at an intermediate velocity (0.005 m/s) from those selected to represent the boundary-lubrication regime. It is obvious that for steel surfaces, higher-viscosity oils provided a more stable friction. In particular, the highest-viscosity oil shows extremely smooth and stable friction behavior (Fig. 6a). This indicates that no film breakdowns occurred, most probably providing quite reliable surface protection, which was not the case at lower viscosities, where significant variations in the friction can be observed at all stages of the test. The worn surfaces clearly support this explanation, since for PAO4 (ISO VG 18) the oil surfaces experience adhesive wear and associated abrasive scratches, while for PAO8



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 5 Comparison of ball wear in steel/steel and DLC/DLC contacts under non-lubricated and lubricated (PAO4) conditions as a function of the number of cycles at a sliding velocity of 0.04 m/s



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 6 Evolution of coefficient of friction for (a) steel/steel and (b) DLC/DLC contacts under lubricated conditions with three different-viscosity oils (PAO4 – ISO VG 18, PAO6 – ISO VG 30 and PAO8 – ISO VG 46) as a function of the number of cycles at a sliding velocity of 0.005 m/s

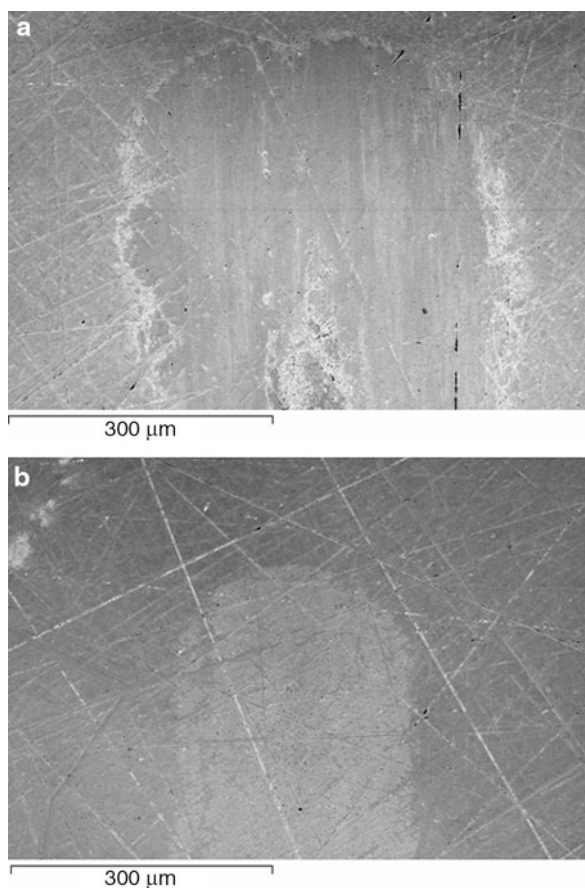
(ISO VG 46) oil, there is almost no visible damage, except some smoothing, on the worn surface (Fig. 7). This behavior thus supports the beneficial effect of oil with a higher viscosity in the boundary regime when having steel/steel contacts, both for wear and friction.

In contrast, in the DLC/DLC contacts, the friction increased when using higher-viscosity oils (Fig. 6b). This is in agreement with the previously shown results (Fig. 4), confirming that – compared with non-lubricated conditions – the friction increases due to the presence of the oil, and with higher viscosity this is even more pronounced. However, at lower viscosity (and lower friction), the friction is less stable and this suggests more interactions between the surfaces (Fig. 6b). However, since DLC surfaces do not experience adhesion, their wear is lower than

in the steel contacts and thus also the friction does not show such a dramatic variation (until the coating is at the surface). However, more detailed analyses presented in the AFM figures of the worn surfaces tested with low- (ISO VG 18) and high-viscosity (ISO VG 46) oils support this suggestion, since much more smoothing, and thus direct contacts between the asperities, can be observed with the low-viscosity oil (Fig. 8). Therefore, exactly as we found initially, the base oil protects the DLC surfaces, and even more if the viscosity is high, but at the same time this increases the DLC friction.

Base-Oil Adsorption at the DLC Coatings

From the above it is clear that the base oil, as well as its rheological properties (viscosity or chain length), clearly



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 7 SEM images of worn steel surfaces tested at a sliding velocity of 0.04 m/s under lubricated conditions with (a) low-viscosity oil (PAO4 – ISO VG 18) and (b) high-viscosity oil (PAO8 – ISO VG 46)

affects the behavior of the lubricated steel/steel and DLC/DLC contacts, although there were no additives in the oil – it was simply “pure,” non-polar, polyalphaolefin oil. This suggests that the oil adsorbs at the surface; however, the oil could also play the “passive” role of a liquid being entrapped in the contact at a high contact pressure, so causing the difference.

In order to investigate these possibilities, quasi-static experiments at very low velocities in the range of $\mu\text{m/s}$ in order to eliminate any effects of velocity and oil entrapment have been additionally performed. With these experiments, the intention was to allow the oil being removed from the real-contact-area asperities by high contact stresses (normal and shear) during asperity collision, and also to eliminate the (viscous) resistance of the oil that is “entrapped” within the contact, since at such a low speed,

the oil could easily migrate through and/or escape from the contact, thus releasing any oil-pressure resistance and not carrying any load. Consequently, if there is no oil adsorption at the surfaces, the oil would play only a minor role in the asperity contacts where most of the load is carried, thus enabling “dry asperity contacts,” and in the case of DLC, the inherent low-friction behavior of these coatings should be re-established.

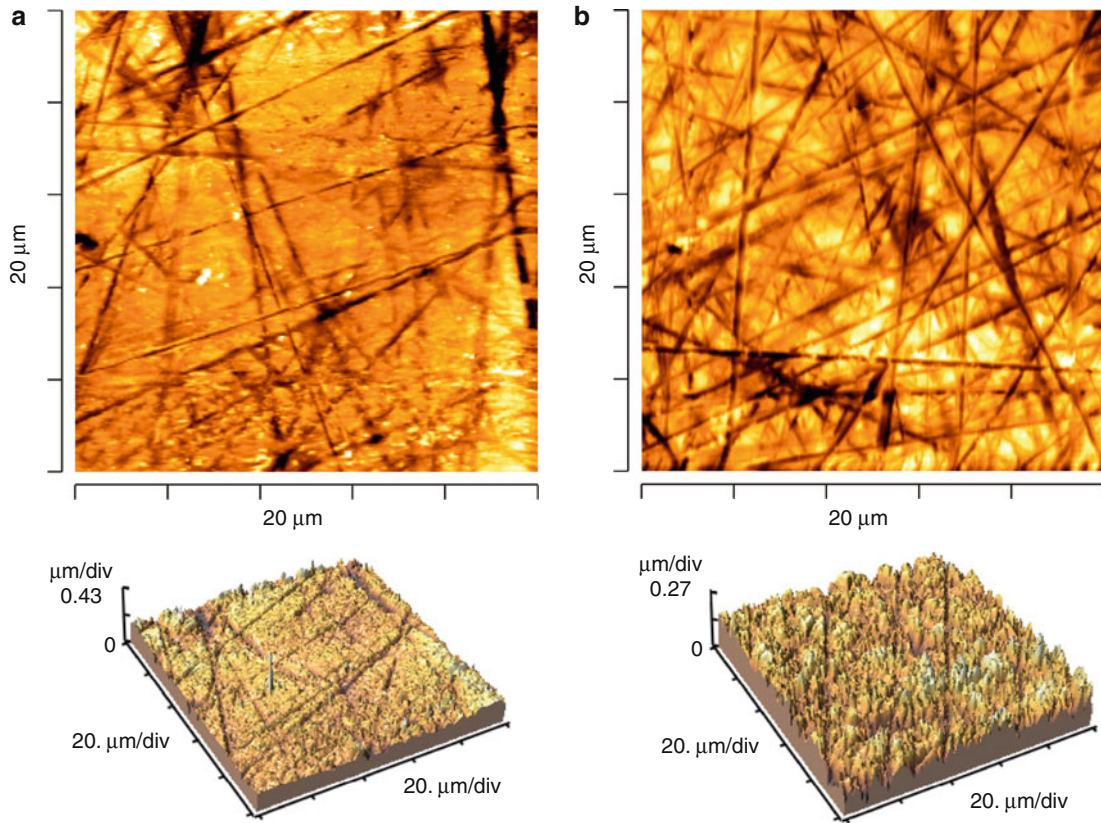
Figure 9a shows the friction of the quasi-static experiments for the steel/steel contacts. The non-lubricated contacts experienced the highest coefficient of friction of 0.25, but this was significantly lower than at higher speeds (0.65–0.75), indicating the reduced effect of the asperity impact energy. The use of the oil (of any viscosity) always reduces the friction, which is another expected result, as a reduction of the adhesive component of friction is well known for steels. In agreement with this, higher-viscosity oils provide a thicker adsorption layer and thus the steel-steel adhesion, which is the key component of the friction in these contacts, is even more reduced.

In contrast, the opposite behavior is observed for the DLC/DLC contacts. The same as with higher velocities (Fig. 1), the lowest friction was measured in the non-lubricated tests, again suggesting the inherent low-friction properties of DLC. Adding low-viscosity oil (PAO4-ISO VG 18) to the DLC/DLC contacts increased the friction by more than twofold, which is a significant difference. A further increase in the oil viscosity resulted in an even higher friction increase, compared to non-lubricated tests, up to 4.5-fold.

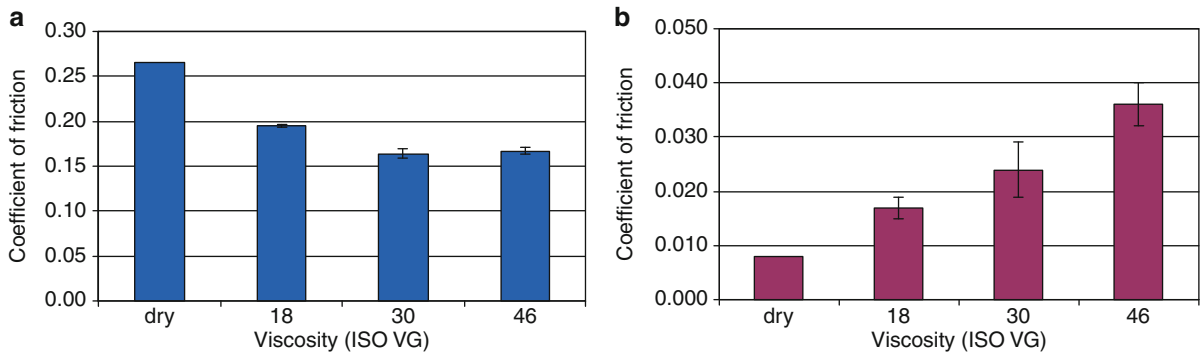
Comparing this DLC behavior with the results for the steel contacts, where a higher viscosity with a better/thicker adsorbed oil layer clearly tends to mainly reduce adhesion – not increase the resistance to motion – supports the hypothesis that the oil was indeed adsorbed at the DLC surface and provided better wear protection, but at the same time resulted in a higher friction due to the asperities sliding over each other on thicker and more viscous films, which by definition causes higher viscous resistance to sliding.

Discussion

Although understanding of the boundary lubrication of DLC coatings has progressed significantly over recent years, it is almost totally unclear – in terms of details – as to what is the actual role of the base oil in the DLC contacts, and particularly, what are the actual contributions of the different “components” in such lubrication, i.e., the role of the base oils with their physical properties, as well as the adsorption behavior, and what is the role of the DLC coating itself? In this work there is an attempt to



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 8 AFM images of worn DLC surfaces tested at a sliding velocity of 0.04 m/s under lubricated conditions with (a) low-viscosity oil (PAO4 – ISO VG 18) and (b) high-viscosity oil (PAO8 – ISO VG 46)



Solid-Liquid Bi-phase Lubricating Coatings, Fig. 9 Coefficient of friction for (a) steel/steel and (b) DLC/DLC contacts under non-lubricated and lubricated conditions with three different viscosity oils (PAO4 – ISO VG 18, PAO6 – ISO VG 30 and PAO8 – ISO VG 46) in quasi-static experiments with velocity 500 $\mu\text{m/s}$

answer some of these questions by using very low velocities relative to the Stribeck-curve parameter, i.e., far into the boundary regime. Since the tribological in situ adsorption effects, especially the physical-based effects, are

difficult to test and analyze with ex situ techniques, the concept of comparing the behavior of DLC coatings with the well-known boundary-lubrication effects of steels under the same conditions was used. Some of the

experiments were performed at very low velocities, i.e., under quasi-static conditions in the $\mu\text{m/s}$ range, to be able to separate and evaluate only the asperity-contacts properties instead of the whole contact.

Non-lubricated sliding caused very high wear and friction of the steel/steel and the DLC/DLC contacts (Figs. 1 and 2). The wear of the steel/steel contacts is well known to be high at large stresses around 1 GPa due to the high adhesion and asperity deformation and the friction is also high because of the strong metallic bonds. In the case of DLC contacts with relatively rough surfaces and a load carried over only a small number of asperities (dry contacts under high load leading to high stress), the micro-fracture process initiates the surface damage and the wear or fracture could progress quite quickly due to low coating-fracture toughness or the low coating-substrate adhesion. These severe conditions were also measured in non-lubricated experiments, clearly pointing out the poor tribological behavior under non-lubricated conditions for both the steel/steel and the DLC/DLC contacts (Figs. 1, 2, 3).

When the base oil was used in the experiments, the friction and the wear of the steel/steel contacts were drastically reduced (Figs. 4 and 5), which was expected. This occurs mainly due to the reduction of the strong adhesion and material transfer between the steel surfaces due to the suppression of the formation of metallic bonds. For higher-viscosity oils, the stronger film protects the surfaces even better and the wear is further reduced (Fig. 7).

The behavior of the lubricated DLC/DLC contacts is more complex. There was, similar to the steel/steel contacts, a significant reduction in the observed wear, where instead of an always-occurring wear through or delamination under dry conditions, the wear became almost negligible (Fig. 5). There are at least two non-chemical, base-oil effects that could explain the DLC wear reduction: (1) One very important effect is the reduction of asperity stresses, which are more uniformly distributed over the entire contact under lubricated contacts (hydrostatic pressure), instead of occurring only at the asperity contacts. This significantly reduces the potential for an initial fracture or damage to the coating surface and the subsequent progress of spallation, as can be seen in Fig. 3. (2) Another influence that affects the reduction in wear is the oil's viscosity. Higher-viscosity oils provide thicker films and thus reduce the possibilities for high-energy asperity impacts, and, consequently, surface damage and wear. The wear results with the higher-viscosity oils strongly support this suggestion (Fig. 8).

However, in contrast to the steel/steel contacts and the expected behavior based on the DLC/DLC wear results,

the friction in the boundary-lubricated DLC/DLC contacts increases instead of decreasing. This appears surprising at first. Nevertheless, it is well known that DLC coatings are low-friction, low-adhesion, and anti-stiction coatings under dry conditions (Erdemir and Donnet 2006; Sanchez-Lopez et al. 2003; Erdemir 2004). Therefore, their low friction under dry conditions is not unexpected. However, when oil is added to the contact, this obviously significantly disturbs these inherent, beneficial, DLC, low-friction properties. High-viscosity oils in a DLC/DLC contact obviously exhibit more of an additional friction shear resistance to be overcome during the sliding due to the thicker oil films, rather than reducing the adhesion component of the friction (which is the dominant beneficial effect with steels). This is because the adhesion does not occur with these DLC coatings, and thus higher-viscosity oils only lead to an even higher friction (Fig. 6) than in dry conditions, although with lower wear (Fig. 8).

Therefore, the base oil in boundary-lubricated DLC/DLC contacts is very important and has a beneficial effect on the DLC coatings' life (endurance), eliminating coating wear and delamination, but increasing friction compared with non-lubricated DLC/DLC contacts (until there is an intact coating at the surface). Moreover, it should be stressed that both the lubricated friction and wear of the DLC/DLC contacts were always lower compared with the steel/steel contacts. So, in spite of some friction increase, this should not be considered as a negative – but as an overall positive tribological effect.

However, based on these results, it was not clear whether there was any base-oil adsorption at the DLC surface or whether the oil was simply a passive viscous fluid “entrapped” in the contact and/or, how strongly such a film – whatever its bonding may be – protects the surfaces and affects the friction of the DLC coatings.

With quasi-static experiments (Fig. 9), the intent was to allow the oil to be removed from the highly stressed, real-contact-area asperities, and also to eliminate the resistance of the oil “entrapped” within the contact, since at such a low speed, the oil could easily migrate and/or escape from the contact and release any oil-pressure resistance, as well as not carrying any load under such conditions. The results confirmed that in steel/steel contacts, the oil (and with higher viscosity even more so), which is known to adsorb at the steel surface, protects the surface through the reduction of adhesion at these high loads (Fig. 9a), rather than increasing the viscous resistance. This is also an indication that the concept of quasi-static experiments with an extremely low λ value (0.003–0.009) was properly selected to analyze the asperity

spot-to-spot contacts behavior, not the viscous resistance of the oil entrapped in the contact.

Consequently, following the same justification, it appears that the friction in the DLC/DLC contacts increases by 2–4.5 times compared with the dry contacts (Fig. 9b), clearly showing that the oil film cannot be removed from the highly stressed contact asperities at the real-contact area, but remains in the asperity interface during the sliding over each other. Moreover, the friction increase use of lubricated contacts and even more so with higher-viscosity oils, was extremely high, indicating that this must be a consequence of the sliding of direct contacts where the majority of the load is carried, not of the surrounding oil (i.e., viscous resistance). This suggests that the oil must have been adsorbed at the DLC surface, otherwise the low-friction “dry-type” DLC/DLC sliding could be re-established. Some further experiments with different test conditions, lower viscosities, and different chain lengths and chemistries were also performed, and they all support the above suggestion. These findings are partially reported elsewhere (Velkavrh and Kalin *in press*), since further discussion on all these details would exceed the scope of this work. In spite of this, other more direct evidence about the base-oil adsorption with respect to DLC should further support the results presented here.

Conclusions

1. When (polyalphaolefin) base oil is used in steel/steel contacts, the friction and wear are drastically reduced, while in DLC/DLC contacts, only the wear is (significantly) reduced. However, in contrast, the friction is increased compared with the non-lubricated conditions.
2. The effect of DLC wear reduction by a nonchemical base oil is explained with two phenomena: (1) a more uniform stress distribution over the entire contact that releases high asperity stresses under dry conditions, and (2) higher-viscosity base oils result in thicker lubricating films that protect the DLC coatings better than the lower-viscosity oils.
3. Base oils disturb the well-known, inherent, low-friction property of DLC coatings due to the additional viscous shear required during the sliding over the asperities – but at the same time the oil does not reduce the adhesion component of friction, because the adhesion is not occurring in these coatings as it does with steel.
4. In spite of the increased friction of DLC/DLC contacts under base-oil boundary lubrication, both the friction and the wear were lower compared with steel/steel contacts under these conditions.

5. With specific low-speed, quasi-static experiments, designed to analyze the effect of asperity spot-to-spot contacts, it was found that the oil film cannot be removed from these contacts, even during very high contact pressures with normal and shear stresses, once the oil is introduced, suggesting that the oil must have been adsorbed at the DLC surfaces, rather than acting as a passive fluid film.

Key Applications

Automotive transmission systems, gears, bearings, nozzles, sliding and rolling surfaces, cutting and forming tools, medical applications.

Cross-References

- [Bonding at Surfaces/Interfaces](#)
- [Diamond-Like Carbon Coatings](#)
- [Friction Coefficient](#)
- [Lubricant Viscosity](#)

References

- M.I. De Barros'Bouchet, J.M. Martin, T. Le-Mogne, B. Vacher, Boundary lubrication mechanisms of carbon coatings by MoDTC and ZDDP additives. *Tribol. Int.* **38**, 257–264 (2005)
- S. Equey, S. Roos, U. Mueller, R. Hauert, N.D. Spencer, R. Crockett, Tribofilm formation from ZnDTP on diamond-like carbon. *Wear* **264**, 316–321 (2007)
- A. Erdemir, Diamond-like carbon films, in *Tribology of Mechanical Systems: A Guide to Present and Future Technologies*, ed. by J. Vizintin, M. Kalin, K. Dohda, S. Jahanmir (ASME Press, New York, 2004), pp. 139–156
- A. Erdemir, C. Donnet, Tribology of diamond-like carbon films: recent progress and future prospects. *J. Phys. D: Appl. Phys.* **39**(18), 311–327 (2006)
- M. Kalin, E. Roman, J. Vizintin, The effect of temperature on the tribological mechanisms and reactivity of hydrogenated, amorphous diamond-like carbon coatings under oil-lubricated conditions. *Thin Solid Films* **515**, 3644–3652 (2007)
- M. Kalin, I. Velkavrh, J. Vizintin, Review of boundary lubrication mechanisms of DLC coatings used in mechanical applications. *Meccanica* **43**, 623–637 (2008)
- M. Kalin, I. Velkavrh, J. Vizintin, The Stribeck curve and lubrication design for non-fully wetted surfaces. *Wear* **267**, 1232–1240 (2009)
- M. Kalin, E. Roman, L. Ožbolt, J. Vizintin, Metal-doped (Ti, WC) diamond-like-carbon coatings: reactions with extreme-pressure oil additives under tribological and static conditions. *Thin Solid Films* **518**, 4336–4344 (2010)
- A. Matthews, A. Leyland, K. Holmberg, H. Ronkainen, Design aspects for advanced tribological surface coatings. *Surf. Coat. Technol.* **100–101**, 1–6 (1998)
- R.M. Mortier, S.T. Orszulik, *Chemistry and Technology of Lubricants*, 2nd edn. (Blackie Academic Professional, Glasgow, 1993)
- A. Neville, A. Morina, T. Haque, M. Voong, Compatibility between tribological surfaces and lubricant additives – how friction and wear reduction can be controlled by surface/lube synergies. *Tribol. Int.* **40**(10–12), 1680–1695 (2007)

- L. Ožbolt, M. Kalin, J. Vižintin, J. Jelenc, The adsorption of polar molecules on DLC coatings, in *Proceedings of the ECOTRIB 2009*, Pisa, 2009
- J.C. Sanchez-Lopez, A. Erdemir, C. Donnet, T.C. Rojas, *Surf. Coat. Technol.* **163–164**, 444 (2003)
- I. Velkavrh, M. Kalin, J. Vižintin, The performance and mechanisms of DLC-coated surfaces in contact with steel in boundary-lubrication conditions – a review. *Stroj. Vestn. J. Mech. Eng.* **54**(3), 189–206 (2008)
- I. Velkavrh, M. Kalin, J. Vižintin, The influence of viscosity on the friction in lubricated DLC contacts at various sliding velocities. *Tribol. Int.* **42**, 1752–1757 (2009)
- I. Velkavrh, M. Kalin, Comparison of the effects of the lubricant-molecule chain length and the viscosity on the friction and wear of diamond-like-carbon coatings and steel (in press)

Solvation Force (Hydration Force Is One Type of Solvation Force)

- [Hydration Force](#)

Sound Velocity

- [Temperature and Pressure Dependence of Density and Thermal Conductivity of Liquids](#)

Soybean Oil

- [Natural Oils as Lubricants](#)

Space Tribology

- [Tribiochemistry in Space Lubrication](#)

Spacing Loss

- [Head Disk Interface for Patterned Media](#)

Spalling

- [Rolling Bearing Fatigue Life, Effect of Profiles, Effect of Surface Roughness, Effect of Residual Stress](#)

Spatial Roughness Parameterse

- [Frequency and Autocorrelation Function](#)

Specific Sliding (Slide/Sweep) Ratio

- [Gear Sliding](#)

Spectral Analysis of Contact Problems

- [FFT-Based Methods for Contact Mechanics](#)

S-Phase

- [Low Temperature Carburization](#)

Spherical Fast Fourier Transform (SFFT) for Contact in Spherical/Aspheric Bearings

- [Contact Mechanics for Spherical/Aspheric Bearing](#)

Spherical Fast Fourier Transform (SFFT) for Elasticity of Spherical Bearings

- [Elasticity Theory for Spherical Bearings](#)

Spherical Grid Data Model (SGDM) for Contact in Spherical/Aspheric Bearings

- [Contact Mechanics for Spherical/Aspheric Bearing](#)

Spherical Grid Data Model (SGDM) for Elasticity of Spherical Bearings

- [Elasticity Theory for Spherical Bearings](#)

Spherical Inverse Filter Method (SIF)

- [Contact Mechanics for Spherical/Aspheric Bearing](#)

Spherical Multi-Grid Technique (SMG)

- [Elasticity Theory for Spherical Bearings](#)

Spherical Race Bearings

- [Self-Aligning Bearings](#)

Spherical Roller Bearings

- [Self-Aligning Bearings](#)

Spiral Groove Bearings

- [Gas Bearings with Narrow Inclined Grooves](#)

Spiral Grooved Conical Bearing

- [Gas Bearings with Narrow Inclined Grooves](#)

Spiral Grooved Opposed-Hemispherical Bearing

- [Gas Bearings with Narrow Inclined Grooves](#)

Spiroid® and Helicon® Gearing

DuWAYNE PAUL

Spiroid® – A Business Unit of ITW Heartland, Alexandria, MN, USA

Synonyms

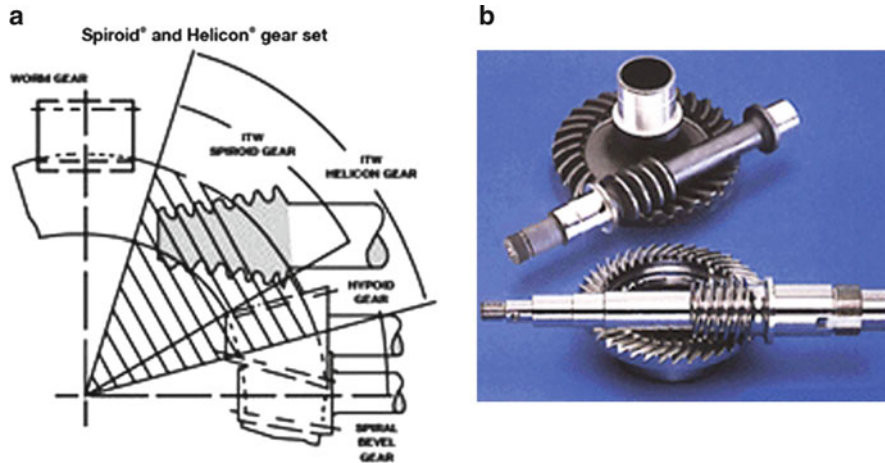
Skew axis gearing with tapered pinion; Skew axis gearing without tapered pinion

Definition

The Spiroid family of gears consists of two trademarked brands, Spiroid® and Helicon®. This type of gearing, often referred to as “skew axis gearing,” operates on nonintersecting and nonparallel axes. It is designed and produced using software, tooling, and methods developed by Illinois Tool Works, who owns and maintains the IP. It is exclusively designed for right-angle power transmission where high amounts of torque are required from a small gearbox envelope. A variety of materials can be used, including steel, brass, aluminum bronze, forgings, powder metal, and plastic (either molded or cut). Typical features include high torque capability, high stiffness, quiet running, compact and lightweight, and easy to assemble. Ratios are available ranging from 3:1 to more than 400:1.

Figure 1a shows the relationship of Spiroid® and Helicon® to traditional right-angle power transmission methods (bevel gearing and worm gearing). **Figure 1b** shows examples of Spiroid® and Helicon® gear sets.

The offset from the gear centerline allows the Spiroid® and Helicon® gearing to maintain more tooth surface to be in contact at any one time, thus increasing the contact ratio. By increasing the contact ratio, higher amounts of torque capacity are generated and motion transmission is smoother.



Spiroid® and Helicon® Gearing, Fig. 1 Spiroid® and Helicon® gearing. (a) Spiroid® and Helicon® compared with bevel and worm gear sets. (b) Photograph of Spiroid® and Helicon® gear sets

Scientific Fundamentals

Key Characteristics

- Simple and positive backlash control – in some applications, backlash can be maintained at zero
- Available as either right hand or left hand
- Can be driven either direction
- Ability to adjust mounting distance based on assembly variation
- Very high accuracy for precise positioning or indexing and constant velocity
- Very high reduction ratios possible
- Superior shock strength
- Very high torque capability in relationship to size of gear set
- Quiet operation
- Versatility of material that can be used
- Compact and lightweight
- High stiffness
- Design flexibility – gear sets are designed to meet the application requirements, rather than developing the application requirements to meet the gearing method

Contact Characteristics

All gears can, at best, obtain only line contact under no load conditions, at any instantaneous position of mesh. The total length of line contact is one criterion of gear load carrying capabilities. Another is contact line movement during the engagement cycle. The contact line should not be stationary. It should sweep the entire available tooth surface. Movement or sweep of the contact line brings freshly lubricated and cool areas into mesh. This applies

to all gears. Other factors, such as relative curvature of contacting surfaces and inclination of the contact lines relative to sliding velocity, must also be considered. With Spiroid® and Helicon® pinions, the line of instantaneous contact is an almost radial line on each convolution of the pinion thread. It is almost perpendicular to the sliding velocity, and results in a full sweep of contact on the pinion and gear.

By contrast, the instantaneous line of contact for worm gears is only slightly inclined to the direction of sliding velocity. This results in a narrow band of contact on the worm thread.

Gear Rotation

Depending on the hand of the pinion thread, a Spiroid® or Helicon® gear set is called either right hand or left hand. Either hand can be positioned with the gear in any of four different gear quadrants, allowing for flexibility in the gearbox design. The relative directions of rotation and preferred directions of rotation allow for maximum bearing life.

Indexing and Rotational Accuracy

Indexing accuracy of gears is often given in minutes or seconds of arc. Other applications demand that the rotational speed of the gear not vary more than a fixed percentage of the average speed.

Like all gears, Spiroid® and Helicon® gears inherit their initial accuracy from the machines that produce them. However, Spiroid® and Helicon® gears have features that minimize the effect of manufacturing deviations. For example, the large number of pinion teeth in

simultaneous contact with the gear teeth helps average out some of the deviations existing on the individual teeth.

Gear tooth and pinion thread runouts cause angular velocity fluctuation, the magnitude of which is effected by the pressure angle. The greater the pressure angle, the greater the fluctuation in angular velocity. Therefore, the low pressure angle side should be used as the driving side in applications requiring low angular velocity fluctuations.

Pressure angles as low as 5° are attainable in many reduction ratios for both Spiroid® and Helicon® gears.

Multiple Tooth Contact

Spiroid® and Helicon® gears have many teeth in simultaneous contact. This, coupled with the fact that each pinion tooth contacts its mating gear tooth along a line perpendicular to the sliding velocity, leads to a number of important advantages.

The number of teeth in contact depends on the number of teeth in the gear member. Generally, 10% of the gear teeth are in simultaneous contact. However, even in the low ratio range, there is two to three times the number of teeth in contact, as compared with worm gears. On higher ratios, there are many times more.

For any gear of a given diameter, a higher ratio means more teeth, which means a finer pitch. Since contact between the Spiroid® and Helicon® pinion and gear extends over the entire length of the pinion, a finer pitch or shorter lead gives a proportionately increased number of teeth in simultaneous contact, and thus, very little sacrifice in capacity (higher torque capability from smaller gears). With worm gears, for instance, the upper ratio limit is usually about 80:1. After selecting a pitch sufficient to carry the required load (bearing in mind that, at most, the worm set has two teeth in contact), the gear becomes too large. Consequently, in the case of worm gearing, higher ratios are usually handled by multiple reductions.

Spiroid® and Helicon® gears, however, do not have an upper limitation. For a given gear diameter, a higher ratio means a shorter lead and finer pitch, and with more teeth in contact. As a result, single reduction ratios of up to 400:1 are possible.

High Side and Low Side

Spiroid® and Helicon® tooth pressure angles are not symmetrical. However, they operate at about the same efficiency and load rating in either direction of pinion rotation, even though direction and magnitude of the forces on the teeth are different. While the resultant tooth forces are largely in the axial direction of the pinion (particularly in the higher ratio range), they also have

components working in the radial direction of the pinion. This tends to separate the pinion and gear teeth. The radial component is greater when the high pressure angle of the tooth does the driving.

Materials

The more favorable contact conditions and lower sliding between Spiroid® and Helicon® gear members permits the use of steel grades that can be hardened, for both gears and pinions. This, plus the ideal lubricating film formation, and the movement of the contact lines, makes the use of hardened steel-on-steel an ideal material for both Spiroid® and Helicon® gearsets. In worm gears, by contrast, the usual selection is a hardened steel worm meshing with a bronze gear.

Quietness

Gear accuracy is the most important factor in obtaining maximum quietness and minimum vibration. However, at high pitch line velocities, even small inaccuracies can produce a pronounced noise, particularly if they occur at regular frequencies. Therefore, to achieve quiet gears, accuracy is combined with modifications in the profile and lead of the gear teeth, permitting them to engage gradually. It was previously pointed out that the contact lines on Spiroid® and Helicon® pinions are nearly radial, and sweep the entire length of the teeth. At such great tooth length, very effective modifications can be made at the entering and leaving side, so the pinion teeth “cam” into and out of the action gently. The commonly used term for these modifications is *crowning*. Incorporated in either the pinion of the gear, and in conjunction with the large number of teeth in simultaneous engagement, it accounts for the fact the Spiroid® and Helicon® gears are inherently very quiet.

Efficiency

The efficiency of a gear set is a measure of the power lost in the gear mesh, which turns into heat that must be dissipated. Thus, a gear set that is 70% efficient loses 30% of the power input and transmits 70%. The efficiency of Spiroid® and Helicon® gear sets, as well as all other gear types, is a function of the gear set geometry (i.e., the efficiency angles) and coefficient of friction. Spiroid® and Helicon® gear set geometry and ideal lubricating characteristics provide for inherently high efficiencies when compared with worm gears.

As with all other gear types, the higher the gear set ratio, the lower the efficiency. Sometimes, however, the low efficiency can be an asset, particularly when “self-locking” is desired. Self-locking is the inability of the

gear to drive the pinion, or back drive, and is a characteristic common to higher ratios gear sets. Special design considerations can accommodate application requirements for higher dynamic efficiencies or self-locking efficiencies through the custom design versatility of Spiroid® and Helicon® gearing.

Mounting

Spiroid® and Helicon® pinion mountings are more rigid than worm gearing because the bearings can be located very close to the working gear mesh. The cantilever-style mounting of the pinion conforms to accepted bevel gear practice. The cantilever design may be changed to a through shaft design for additional rigidity. Another useful mounting is the short straddle mount, where the through shaft is supported by a needle bearing. This mounting becomes particularly useful where extremely high stiffness is required.

Axial Positioning and Backlash Control

Many modern gear applications demand controllable “near zero” backlash control. For example, applications requiring precise positioning or indexing, or systems with frequent load reversals, benefit from simple backlash adjustment.

Both Spiroid® and Helicon® pinions have threads of constant lead and pressure angle. A Helicon® is completely insensitive to its axial position. Spiroid® pinions are insensitive within the limits of movement necessary for backlash control.

The same is true for Spiroid® and Helicon® gears as far as their axial position is concerned. This feature provides for easy backlash adjustment by merely moving the gear along its axis.

Since the Spiroid® pinion also is adjustable along its axis within a range of positions, there are new opportunities for finer backlash control. No other type of gearing offers this feature in such a simple and direct way.

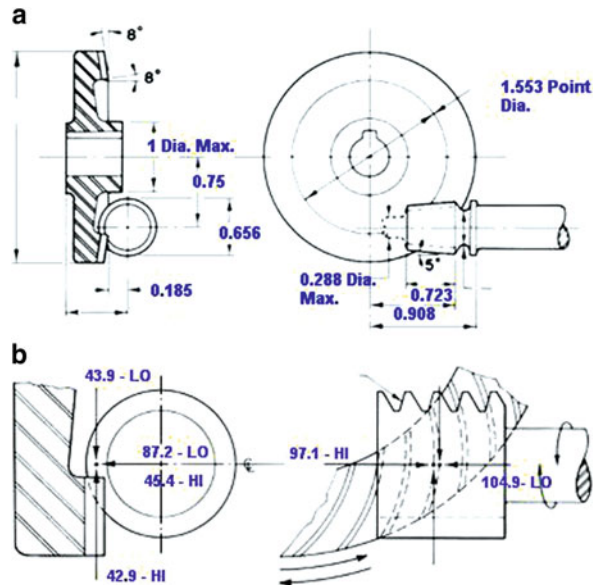
Key Applications

Spiroid® and Helicon® gearing is especially advantageous for applications that require a right-angle gear set and a high torque in a small space, and that must be lightweight. Traditionally, these have been in military and aerospace applications, but it can be used in any application with similar requirements.

Examples of Spiroid® and Helicon® Gearing

Below are examples of the two gearing types addressed here. They are compared with each other with the same parameters. These examples can then be taken and compared with worm gearing or bevel gearing using the same design parameters.

Example 1 – Spiroid® Tooth Style – Reference:



Parameters:

- 2.250 in. gear OD
- Material – steel on steel
- Moderate shock load
- Electric motor input – 24 h per day
- Pinion speed 1,200
- Ratio – 10:1
- Peak torque output – 170 in. pounds

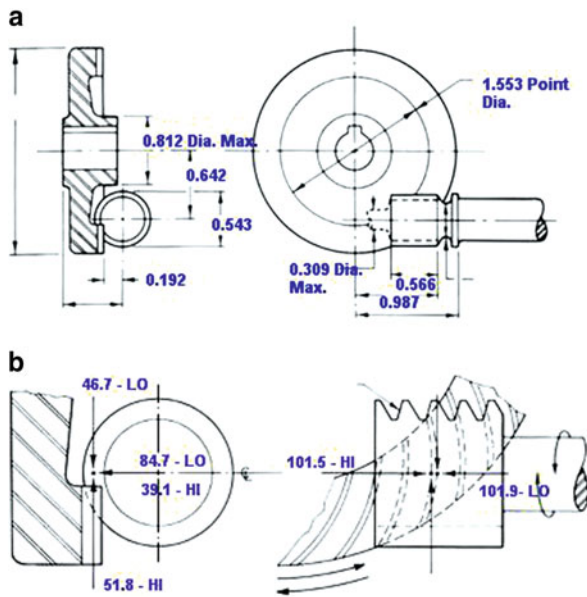
Dimension data		Tooth load data	
		Low side	High side
Pressure angle low – 15	Fx	1.049	0.971
Pressure angle high – 35	Fy	0.454	0.872
	Fz	0.429	0.439

Efficiency = .866

Material Factor = 1.00

Output Capacity = 170

Service Factor = 1.50

Example 2 – Helicon® Tooth Style – Reference:**Parameters:**

- 2.250 in. gear OD
- Material – steel on steel
- Moderate shock load
- Electric motor input – 24 h per day
- Pinion speed 1,200
- Ratio – 10:1
- Peak torque output – 170 in. pounds

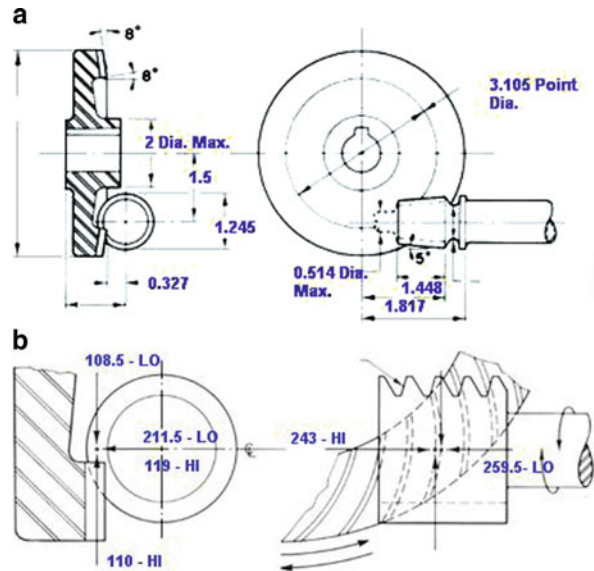
Dimension data	Tooth load data		
		Low side	High side
Pressure angle low – 15	Fx	1.019	1.015
Pressure angle high – 32.5	Fy	0.391	0.847
	Fz	0.518	0.467

Efficiency = .847

Material Factor = 1.00

Output Capacity = 170

Service Factor = 1.50

Example 3 – Spiroid® Tooth Style – Reference:**Parameters:**

- 4.500 in. gear OD
- Material – Steel on Steel
- Moderate shock load
- Electric motor input – 24 h per day
- Pinion speed 1,200
- Ratio – 10:1
- Torque output – 1,133 in. pounds

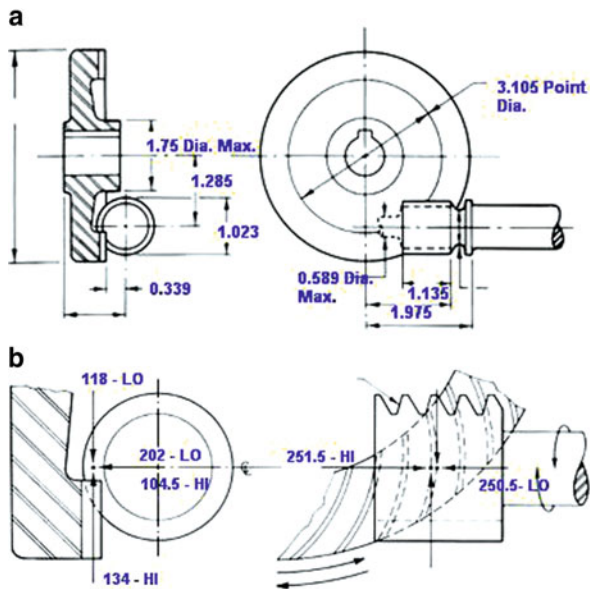
Dimension data	Tooth load data		
		Low side	High side
Pressure angle low – 15	Fx	0.519	0.486
Pressure angle high – 32.5	Fy	0.238	0.423
	Fz	0.238	0.217

Efficiency = .878

Material Factor = 1.00

Output Capacity = 1,133

Service Factor = 1.50

Example 4 – Helicon® Tooth Style – Reference:**Parameters:**

- 4.500 in. gear OD
- Material – Steel on Steel
- Moderate shock load
- Electric motor input – 24 h per day
- Pinion speed 1,200
- Ratio – 10:1
- Torque output – 1,140 in. pounds

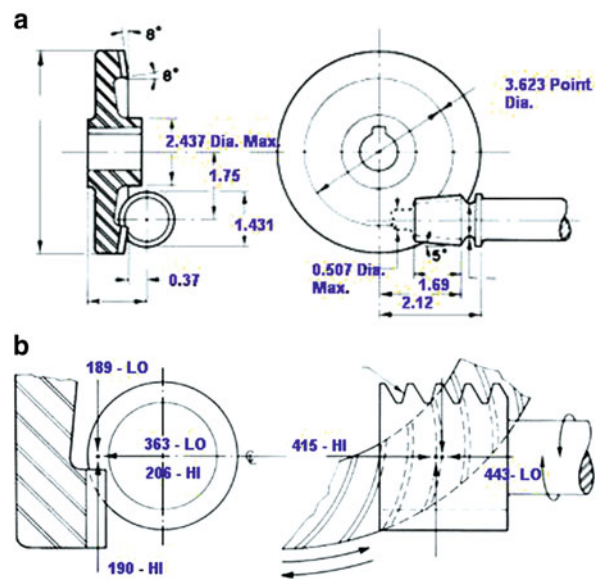
Dimension data	Tooth load data		
		Low side	High side
Pressure angle low – 15	Fx	0.501	0.503
Pressure angle high – 30.0	Fy	0.209	0.404
	Fz	0.268	0.236

Efficiency = .894

Material Factor = 1.00

Output Capacity = 1,140

Service Factor = 1.50

Example 5 – Spiroid® Tooth Style – Reference:**Parameters:**

- 5.250 in. gear OD
- Material – steel on steel
- Moderate shock load
- Electric motor input – 24 h per day
- Pinion speed 1,200
- Ratio – 10:1
- Torque output – 1,698 in. pounds

Dimension data	Tooth load data		
		Low side	High side
Pressure angle low – 15	Fx	0.443	0.415
Pressure angle high – 32.5	Fy	0.206	0.363
	Fz	0.190	0.189

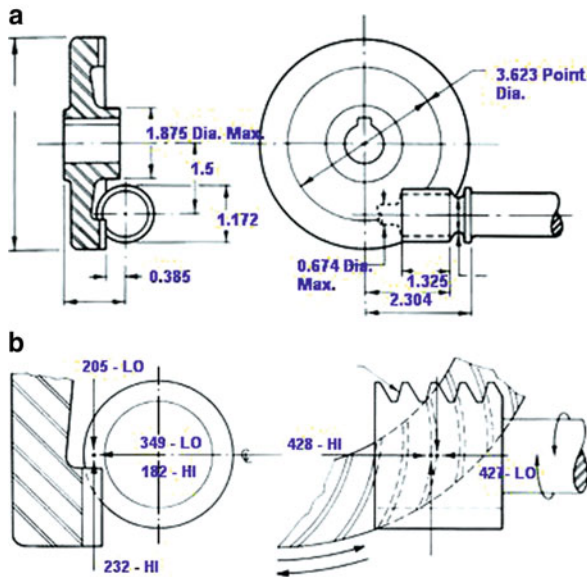
Efficiency = .884

Material Factor = 1.00

Output Capacity = 1,698

Service Factor = 1.50

Example 6 – Helicon[®] Tooth Style – Reference:



Parameters:

- 5.250 in. gear OD
- Material – steel on steel
- Moderate shock load
- Electric motor input – 24 h per day
- Pinion speed 1,200
- Ratio – 10:1
- Torque output – 1,709 in. pounds

Dimension data	Tooth load data		
		Low side	High side
Pressure angle low – 15	Fx	0.427	0.428
Pressure angle high – 30.0	Fy	0.182	0.349
	Fz	0.232	0.205

Efficiency = .90

Material Factor = 1.00

Output Capacity = 1,709

Service Factor = 1.50

Cross-References

- ▶ [Average Reynolds Equations](#)
- ▶ [EHL Film Thickness Behavior](#)
- ▶ [Lubrication Regimes](#)
- ▶ [Navier-Stokes Equation and Applications in Lubrication](#)
- ▶ [Newton's Law of Viscosity, Newtonian and Non-Newtonian Fluids](#)

- ▶ [Reynolds Equation](#)
- ▶ [Rheological Measurement Methods and Equipment](#)
- ▶ [Rheology – Viscosity Index](#)
- ▶ [Temperature and Pressure Dependence of Viscosity](#)
- ▶ [Viscosity Index Additives](#)

References

- F. Bohle, O. Saari. *Spiroid gears – a new development in gearing*. AGMA Paper No. 389.01, 1955
- D.W. Dudley, *Handbook of Practical Gear Design* (CRC Press, Baco Raton, 1994)
- ITW Spiroid, *Spiroid Gearing Design Manual No. 6*, Illinois Tool Works, 1986
- F.L. Litvin, *Development of gear technology and theory of gearing*. NASA Reference Publication 1406, ARL-TR-1500, Biography 3.15, Oliver E. Saari- Inventor at the Illinois Tool Works (ITW) Spiroid Division, 1997
- O.E. Saari. *Skew axis gearing*. U.S. Patent No. 2,954,704, 1960
- O.E. Saari. *Speed-reduction gearing*, U.S. Patent No. 2,696,125, 1954

SPL (Solid Pellet Lubricants)

- ▶ [Carrier-Free Die Casting Lubricants](#)

Spur Gears

ROBERT ERRICHELLO

GEARTECH, Townsend, MT, USA

Synonyms

[Straight gears](#); [Straight-cut gears](#)

Definition

A spur gear has a cylindrical pitch surface and teeth that are parallel to the gear axis. External spur gears have teeth that project outwards, whereas internal spur gears have teeth that project inwards.

Scientific Fundamentals

Contact Ratio

To obtain conjugate action, involute spur gears must have two pairs of teeth simultaneously in contact near the start and end of engagement. Consequently, involute spur gears are typically designed with a transverse contact ratio of at least 1.2, and preferably nearer to 1.6. Equation (1) gives

the percentage time that a single pair of teeth is in contact. For example, a gearset with a contact ratio of 1.6 has a single pair of teeth in contact 25% of the time and two pairs of teeth in contact 75% of the time.

$$\eta_1 = 100(2 - \varepsilon_\alpha)/\varepsilon_\alpha \quad (1)$$

Involute spur gears designed to conventional proportions have a transverse contact ratio less than two ($\varepsilon_\alpha < 2$). High contact ratio (HCR) spur gears can be designed to obtain $\varepsilon_\alpha > 2$ by using a low pressure angle, tall teeth, or both. However, HCR spur gears have high sliding velocity and increased risk of scuffing.

Lines of Contact

If involute spur gears operate on perfectly parallel axes, and the teeth are perfectly accurate and without lead modification or lengthwise crowning, contact between mating teeth occurs along the whole length of the teeth at each instant of rotation and the line of contact is parallel to the gear axis. In practice, some tooth inaccuracy and some misalignment of the axes is inevitable. Therefore, it is common practice to manufacture the teeth with lead modification, crowning, or both to avoid heavy contact at ends of the faces. These modifications must be limited to avoid shortening the contact line excessively.

The total length of the lines of contact on an uncrowned spur gearset equals twice the face width when teeth mesh near the start and end of the engagement in the two-pair zones. Near the pitch point, the teeth mesh in the single-pair zone, and the length of the contact line equals the face width.

Stiffness Variation

If involute spur gears do not have profile modification in the form of tip relief or root relief, the mesh stiffness changes abruptly as the teeth move through the engagement zone, and the number of teeth in contact varies between two-pair and single-pair contact. Mesh stiffness variation causes non-conjugate meshing, which results in noise, vibration, and dynamic tooth forces. Therefore, it is common practice to manufacture the teeth with profile modification in the form of tip relief, root relief, or both to mitigate the stiffness variation. These modifications must be limited to avoid shortening the contact ratio excessively.

Advantages of Spur Gears

The simple geometry of spur gears permits manufacture by many methods including milling, hobbing, shaping, die casting, powder metallurgy, stamping, shearing, roll forming, and electric discharge machining (EDM).

Lapping, burnishing, honing, shaving, grinding, and superfinishing can be used to finish them.

Theoretically, spur gears impose only radial loads on their bearings. In practice, misalignment of the gear axes causes some axial thrust, but the magnitude of the thrust force is usually insignificant. Spur planet (idler) gears can be mounted on a single spherical-roller bearing to achieve self-aligning ability and uniform load distribution along the length of the teeth.

Disadvantages of Spur Gears

Spur gears have the following disadvantages:

- Mesh stiffness variation causes non-conjugate meshing, which results in noise, vibration, and dynamic tooth forces.
- The tooth profiles of spur gears must remain accurate to provide conjugate action. Hertzian fatigue such as micropitting often attacks the dedenda of the teeth and degrades the base pitch spacing. Consequently, spur gears are likely to become non-conjugate if they sustain micropitting.
- If a spur gear fails by bending fatigue, the tooth usually fractures full across the face width when the mesh is in the single-pair zone. The failed tooth usually fractures violently and frequently jams the mesh.
- Spur gears tend to trap oil in roots of teeth. This is a principle reason why spur gears are not used in high-speed applications.
- Spur gears do not distribute lubricant in the lengthwise direction. Therefore, lubricant supply jets must overlap to ensure coverage of the full face width.

Key Applications

Spur gears have wide applications, and are one of the most common forms of gears. However, they are generally used in lower speed applications and applications where gear noise is not a principle consideration.

Cross-References

- [Internal Gears](#)
- [Involute Gear Profiles](#)
- [Rack](#)

References

- ANSI/AGMA 1012-G05, *Gear Nomenclature, Definition of Terms with Symbols* (AGMA, Alexandria, 2005)
- E. Buckingham, *Analytical Mechanics of Gears* (McGraw-Hill, New York, 1949). Republished by Dover, New York, 1963
- J.R. Colbourne, *The Geometry of Involute Gears* (Springer, New York, 1987)

T.W. Khiralla, *On the Geometry of External Involute Spur Gears* (T.W. Khiralla, Studio City, 1976)

J.C. Leming, "High Contact-Ratio (2+) Spur Gears," *Gear Design, Manufacturing and Inspection Manual AE-15* (SAE, Warrendale, 1990)

Sputter-Deposited Molybdenum Disulfide Coatings

► Doped MoS₂ Coatings and Their Tribology

Sputter-Deposited MoS₂ Coatings

► Doped MoS₂ Coatings and Their Tribology

Sputtering ↔ Ionic Bombardment

► Vapor Deposition Coating Technologies (CVD, PACVD, PVD, and Hybrid PVD-CVD) and Their Tribological Application

Sputtering MoS₂-based Coatings

IHSAN EFEOGLU

Department of Mechanical Engineering, Atatürk University, Erzurum, Turkey

Synonyms

Closed-field unbalanced magnetron sputtering (CFUBMS); Coefficient of friction (CoF); Critical load (L_c) and coatings; Gram force (gf); Knoop hardness (KH); Physical vapor deposition (PVD) to synthesize MoS₂; Relative humidity (RH); Room temperature (RT)

Definition

MoS₂ continues to attract attention due to its low friction coefficient. It has been observed that an excellent lubricant property is shown when MoS₂ composite films grown by

sputtering are used in dry air, inert gas, or vacuum environments. Many studies have revealed that the oxidation resistance and endurance of MoS₂ films in atmosphere environments increase when deposited by sputtering with the addition of metal in a multilayer or composite structure. Recently, quasi-amorphous MoS₂ composite coatings with Ti/Nb incorporation were developed, and denser films having higher adhesion and oxidation resistance were obtained via addition of titanium into the structure. MoS₂-Ti/Nb composite films created by PVD-CFUBMS (closed field unbalanced magnetron sputtering) exhibit higher hardness, resistance to atmospheric effects, and higher wear life.

Scientific Fundamentals

Lubrication has been used throughout human history. When humans designed and constructed machines, lubricants were used. Early humans probably used some materials (such as water, ice, wet clay, powdered snow, animal fat) to reduce friction and their experiences yielded valuable results. Since the Industrial Revolution, some slippery solids – graphite, talc, mica, molybdenum disulfide – have been used in machine components. A self-lubricated solid is "any solid material used as a powder or a thin film on a surface to provide protection from damage during relative movement and to reduce friction and wear."

Historical factors and the Industrial Revolution have influenced the lubrication techniques applied to most machines. In particular, dry solid lubrications of sliding surfaces using solid thin films/coatings have become a common method in industrial applications. Today, there is no single coating that can provide both low friction and high wear resistance under severe application conditions (Donet and Erdemir 2004). As a solid lubricant, MoS₂ is well known for its lamellar structure with weak van der Waals bonding between crystal planes. However, weak van der Waals bonding exists in the neighboring planes of S atoms. In general, it is believed that the easy sliding of neighboring planes of S atoms on each other is behind the low friction resistance (Fleischauer 1987). The basal plane of MoS₂ has to be parallel to the substrate surface for ideal friction properties (Weise et al. 1995). MoS₂ oxidizes easily in humid environments. Consequently, the oxidation products (oxidation of MoS₂ into MoO₃ and H₂SO₄) cause an increase of the friction coefficient and a decrease of the wear life, thus creating a corrosive and an abrasive effect on the counterface. Many studies have revealed that the oxidation resistance and endurance of MoS₂ films on multilayer or composite structure surfaces are enhanced by sputtering with additional metal components (Simmonds et al. 1999).

Recently, quasi-amorphous MoS₂ composite coatings with Ti incorporation were developed, resulting in denser films having higher adhesion and oxidation resistance. MoS₂-Ti composite films deposited by CFUBMS exhibit higher hardness resistance to atmospheric effects and longer wear life (Arslan et al. 2004; Efeoglu and Bulbul 2005).

There is no previously published literature for the MoS₂-Nb film. It has been reported that NbS₂ is not a good lubricant itself but that additional Nb atoms intercalated into the structure will result in a change in electron bonding to favor the MoS₂ structure (Jaminson et al. 1978). From this point of view, it would be interesting to look at the friction properties of Nb-doped MoS₂ self-lubricating thin film. Efeoglu and co-workers have been interested in Nb-doped MoS₂ thin film. In their research, the tribological performance of MoS₂-Nb film was evaluated under different atmospheric conditions (humid air, dry nitrogen, distilled water, and oil). The reported key points that the friction and wear properties of air, dry nitrogen, distilled water, and oil were determined using a pin-on-disc tribo-tester. Friction coefficient (CoF) and wear rate were obtained in dry nitrogen as maximum and in oil as minimum. MoS₂-Nb films have exhibited one type of crystallite orientation and characteristic dense structure. Also there is a sign of the humidity-sensitive NbS₂ phase, which displayed better lubricating properties in the humid air than in the dry nitrogen conditions (Efeoglu et al. 2008).

In the last decade, a number of multilayer coating systems have been investigated. The demand for high-performance solid lubricant coatings with low friction coefficient and high wear resistance in severe environments is still increasing. The most common dry solid lubricants are graphite, MoS₂, WS₂, TaS₂, and PTFE, among which the most widely used lamellar compound solid lubricant is MoS₂ with hexagonal and anisotropic crystal layer structure. Having very low friction coefficient and high sensitivity to oxidation associated with easy shear planes of the lamellar structure, MoS₂ suffers from rapid failure in air, which makes it more suitable for use in vacuum environment (Bhushan and Gupta 1991; Weise et al. 1995).

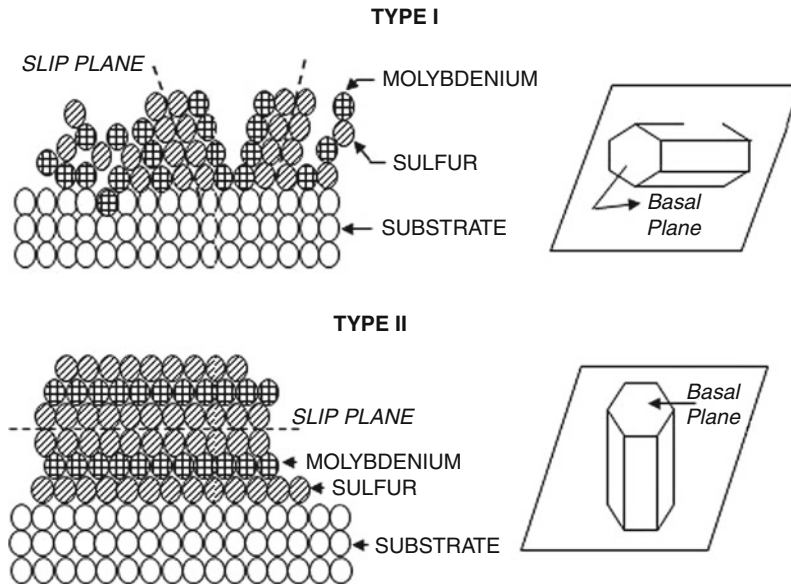
This discussion describes the various single and multilayered composite solid thin films: MoS₂-Ti and MoS₂-Nb as composite single layer and MoS₂:C:TiB₂, TiN:TiAlN-MoS₂:MoS₂-Ti, and TiB₂:TiBN-MoS₂:MoS₂-Ti as composite-graded multilayers currently being used or having potential as solid lubricants. Reviews of solid lubricants are also given in references. Discussions of the nature and influence of surface films, friction, and wear mechanisms are covered in earlier and present work.

Part 1: General Review

MoS₂-based solid thin films display remarkable self-lubrication properties. The commonly used deposition technique today is sputtering. The sputtering technique is well-suited to coat machine elements with uniform films having extraordinary structural, mechanical, tribological, and chemical properties. In this discussion the reviewed literature and deposition method will be based on MoS₂-Ti/Nb and magnetron sputtering.

In space environment, sputtered MoS₂ films display ultra-low friction, but in humid air the friction properties are degraded. For these reasons, sputtered MoS₂ films are primarily used for spacecraft and satellite moving mechanical assemblies and components, which operate under high vacuum, high and low temperatures, and space radiation (Spalvis 1991). Thus, researchers have focused on MoS₂ coatings with metal addition, in which it was determined that MoS₂ films deposited with metal addition such as Ti, Zr, Cr, W, and Nb getter oxygen partly during the wear process and the tribological properties of film improved relatively (Renevier et al. 2000; Arslan et al. 2001; 2008). It is therefore interesting that a number of compounds of transition elements have been studied for solid lubricant use and some of them (NbS₂, MoS₂, WS₂) have been found to be very effective, but no one has yet shown any particular relationship between transition element structure and lubrication performance (Lansdown 1999). The only transition metal dichalcogenides that have shown real promise for lubricant use are the disulfides of W and Nb and the diselenides of Mo, W, and Nb (Renevier et al. 2001). Initially, it has been reported that NbS₂ is not a good lubricant itself but additional Nb atoms intercalated into the structure will result in a change in electron bonding to favor the MoS₂ structure (Jaminson et al. 1978). From this point of view, it would be interesting to look at the friction properties of Nb-doped MoS₂ self-lubricating thin film. There is no previously published literature for the MoS₂-Nb film except the studies on Nb-added MoS₂ by Efeoglu and co-workers (Efeoglu et al. 2008), Arslan and co-workers (Arslan et al. 2005; 2008).

The good lubricating property of MoS₂ coatings is attributed to two parameters (Didziulis et al. 1990): the adhesion of film to substrate and its crystallographic orientation. Two types of orientation, referred to as Type I and II, are known in MoS₂ films grown by sputtering depending on deposition parameters. The basal plane orientation in Type I, allowing penetration of environmental elements, is perpendicular to the substrate surface, as seen in Fig. 1. The basal plane resistant to



Sputtering MoS₂-based Coatings, Fig. 1 Schematic illustration of Types I and II (Taken reference Fleischauer 1987)

environmental attacks is parallel to the substrate in Type II (Fleischauer 1987).

Principles of Dry Self-Lubrication

The main principle of self-lubricated solid thin films is that if a low shear strength material is placed between sliding surfaces in contact, the friction force during the sliding will be reduced (Spalvis 1991). Transition metal sulfides, especially molybdenum sulfide (MoS₂), are scientifically and technologically important materials as self-lubricated thin films. There has been increasing interest in the synthesis of these types of sulfides because of their potential applications in lubrication. MoS₂, h-BN, graphite, and WS₂ are widely used self-lubricated thin films. All these types of thin films have lattice lamella structures with slight tangential loading and these lamella materials also have good load-bearing capacity in sliding and rolling conditions. Graphite has excellent lubricating properties in moisture vapor, and can be used as a lubricant up to approximately 790°C. MoS₂ and WS₂ function well in high vacuum and have higher load-bearing capacity than graphite.

Part 2: Deposition of MoS₂-(Ti/Nb)-Based Coatings

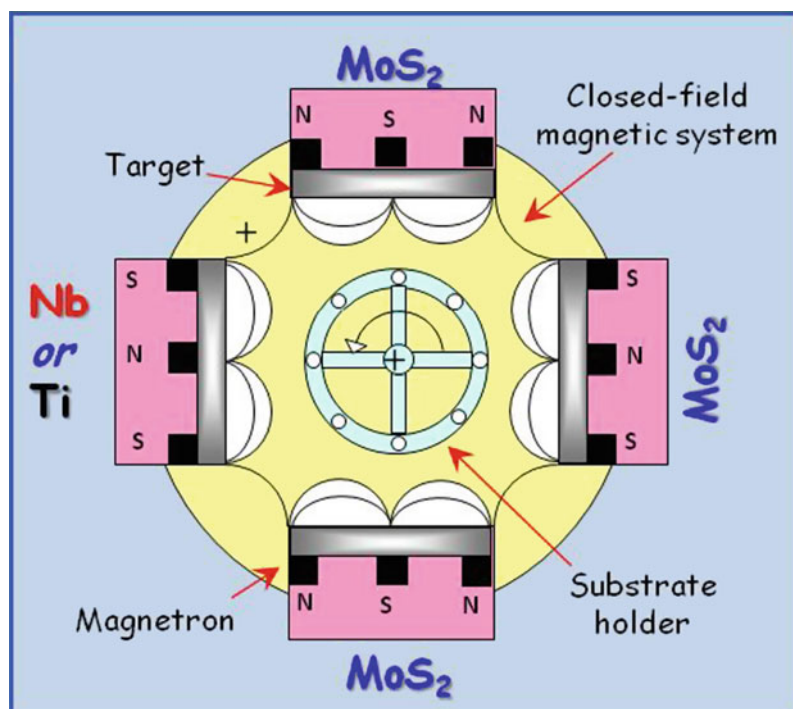
Graded Single- and Multi-Layer Coatings

The demand for high-performance solid lubricant coatings with low friction coefficient and high wear resistance

in severe environments is still increasing. More recently, a Ti addition to MoS₂ has been introduced to the market as MoSTTM by Teer Coatings Ltd. A future direction for the development of low-friction and high wear resistance coatings is a combination of a soft layer coated onto hard layer(s). MoS₂-Ti/Nb composite solid lubricant films were deposited on selected steels by dc-magnetron sputtering (Efeoglu et al. 2005, 2007). Co-sputtering of C and TiB₂ has given the soft MoS₂ an improved tribological property at ambient conditions (Efeoglu 2005) (Fig. 2).

The current trend in modern tribology is to limit or reduce the use of liquid lubricants as much as possible but increase the use of solid lubricant coatings with self-lubricating properties (Donet and Erdemir 2004). It is widely acknowledged that a multilayer film can have many advantages. Load-carrying capacity can be improved. Increased adhesion between the substrate and each individual layer can be obtained possible internal stresses in the nano-superlattice structure. Consequently, multilayered, graded composite coatings have the potential to improve the tribological properties of tools. By alloying with selected elements (Ti, Al, Nb, C, N) and compounds (TiB₂), MoS₂ coatings become applicable to atmospheric conditions. In most cases, the machining is conducted under dry and fluid conditions.

TiN/TiAlN-MoS₂/MoS₂-Ti and Mo:S:C:Ti:B graded multilayers films have been deposited onto D2 tool steels. Deposition parameters have been previously reported (Efeoglu 2005; 2007).



Sputtering MoS₂-based Coatings, Fig. 2 Teer coating closed field unbalanced magnetron coating system (Taken reference Efeoglu et al. 2005, 2007)

Part 3: Properties of MoS₂-Based Self-Lubricated Coatings

Microstructure of Deposited Films

Because pulsed biasing of the substrate during deposition increases the grade of ionization and ion-to-neutral ratio, the composite coatings grow as dense, compact, noncolumnar structures and featureless coating surfaces as shown in Fig. 3a. It was observed that all MoS₂-based coatings deposited by biased-dc magnetron sputtering (CFUBMS) had a very dense microstructure (Fig. 3b).

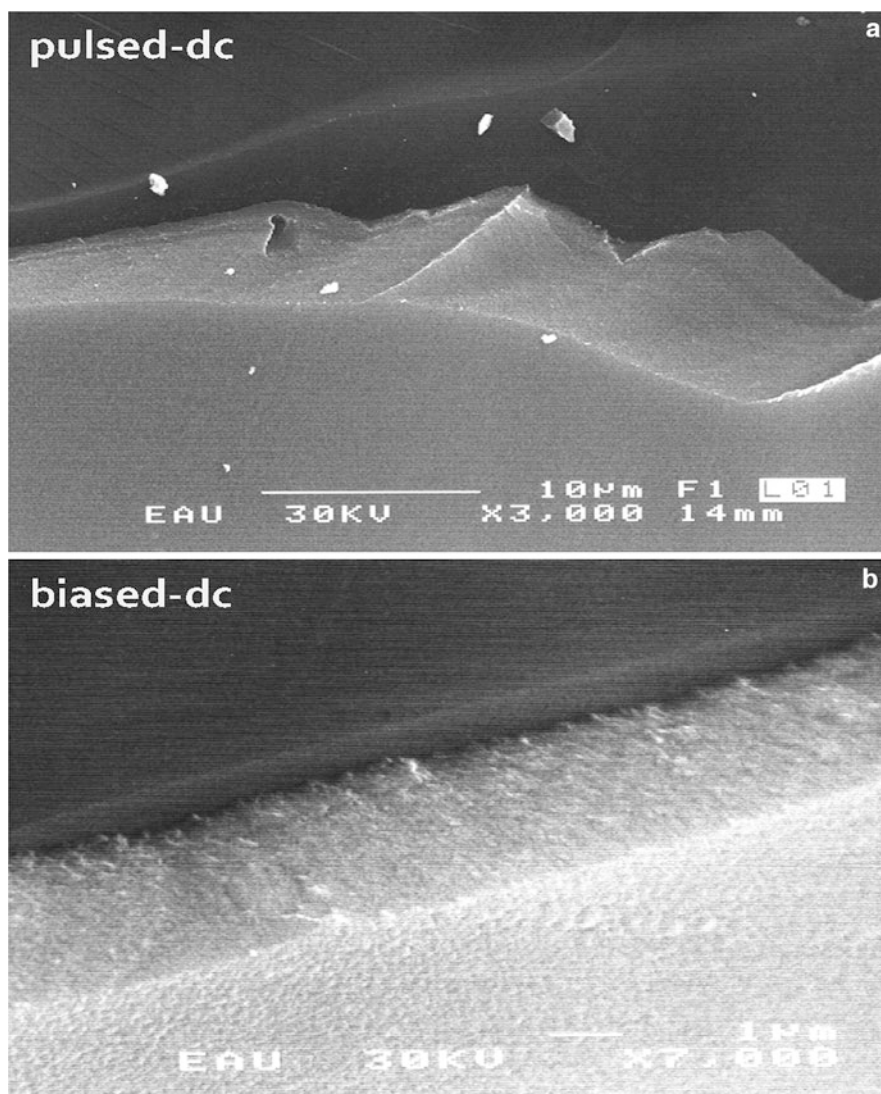
Figure 4 shows an SEM image of the cross-section of a coating in which different structures can be seen. The morphology of the coating is very dense, with little evidence of the columnar morphology of the MoS₂-Ti top layer. The total thickness of the coating was approximately 3 μm (Ti ~ 0.2 μm, TiN ~ 1 μm, TiAlN-MoS₂ ~ 1.5 μm, MoS₂-Ti ~ 0.3 μm). Figure 5 also shows an SEM image of the cross-section of TiB₂:TiBN-MoS₂:MoS₂-Ti multilayer coating.

XRD and EDS Analysis

MoS₂-Ti: Different compositions of MoS₂-Ti composite coatings deposited with pulsed-dc were obtained by

changing deposition parameters (except pulse parameters applied to substrate) (Efeoglu and Bulbul 2005). Elemental composition of the films ranged between 8.54 and 46.6 at.% Ti content and the ratio of sulfur to molybdenum were found to vary between 0.99 and 1.27. The ratio of sulfur to molybdenum was found to be close to unity for run 1 to run 2 processes, but elemental content of the Ti obtained from the run 2 processes had the lowest. XRD analysis indicated all of the MoS₂-Ti composite films coated by pulsed-dc had observable texture except run 3. Several peaks {(0 0 2) and (1 0 0)/(1 0 3)} were evidenced for the composite coatings, indicating that these coatings are at least partly crystalline where MoS₂-Ti are referred to as (2 0 0) oriented coating for runs 1 and 2 (see Fig. 6). In this work, a clear relationship has been found between the oriented basal plane and Ti content in the films. The remarkable results were achieved in terms of coating quality with application of pulsed-dc power to the substrates (see Fig. 3).

MoS₂-Nb: Other work has been done on Nb-added MoS₂ coatings (Efeoglu et al. 2009a). XRD examinations at room temperature indicated that randomly oriented coatings were deposited on the substrate. The reflection at about 2θ = 13° showed the basal plane (002) that was

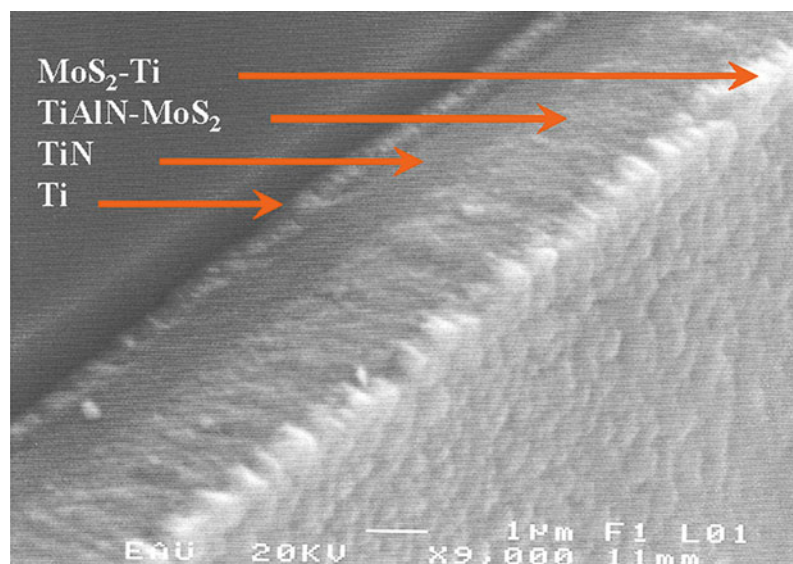


Sputtering MoS₂-based Coatings, Fig. 3 SEM micrograph of fractured cross section of MoS₂-Ti/Nb coatings (Taken reference Efeoglu et al. 2009)

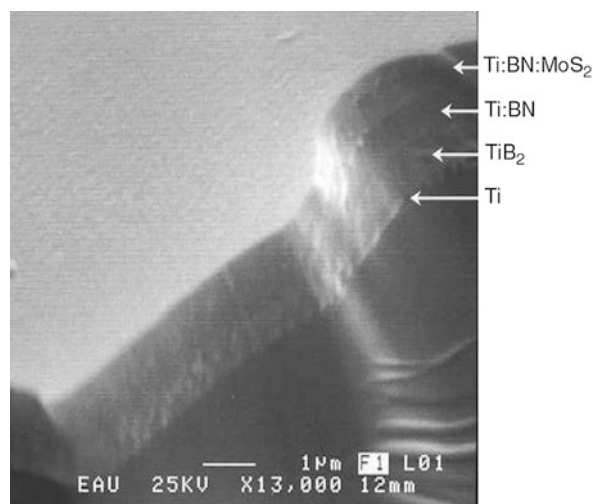
parallel to the substrate except R3, R6, and R8. Figure 7 shows the XRD pattern for MoS₂-Nb sputtered films in which (101) dominant phase, and also MoS₂ (001) (002) and (101) grew. The XRD pattern of the MoS₂-Nb shows a structure in which a single crystal phase (101) grew. There is also a sign from NbS₂ (103). This crystal plane provides a low friction coefficient due to the ease of sliding of the surfaces on each other. On the other hand, the low-order crystals of (101) and (103) peaks showed overlapping of broad peaks from MoS₂ phases doped with Nb atoms.

TiN:TiAlN-MoS₂:MoS₂-Ti: A typical plot of the X-ray diffraction measurement of TiN/TiAlN-MoS₂/MoS₂-Ti multilayer-composite coating is shown in Fig. 8. The MoS₂ phase is obviously nanocrystalline with a strong (002) texture, (100) peak has also been observed. Chemically inert (002) planes parallel to the TiAlN-MoS₂ composite layer, preferred for a tribological application, were observed.

TiB₂:C:MoS₂: XRD analyses were performed on the coatings deposited onto Si-wafer substrates (Efeoglu 2005). Figure 9 shows XRD patterns for co-sputtered Mo:S:C:Ti:



Sputtering MoS₂-based Coatings, Fig. 4 SEM micrograph of fractured cross section of the multilayer-composite thin film (Taken reference Efeoglu 2007)



Sputtering MoS₂-based Coatings, Fig. 5 SEM micrograph of fractured cross section of the multilayer-composite thin film (Unpublished fig: taken Workshop on "Boron using on Defense Industry," SSM Jun 14, 2011, Ankara, Turkey)

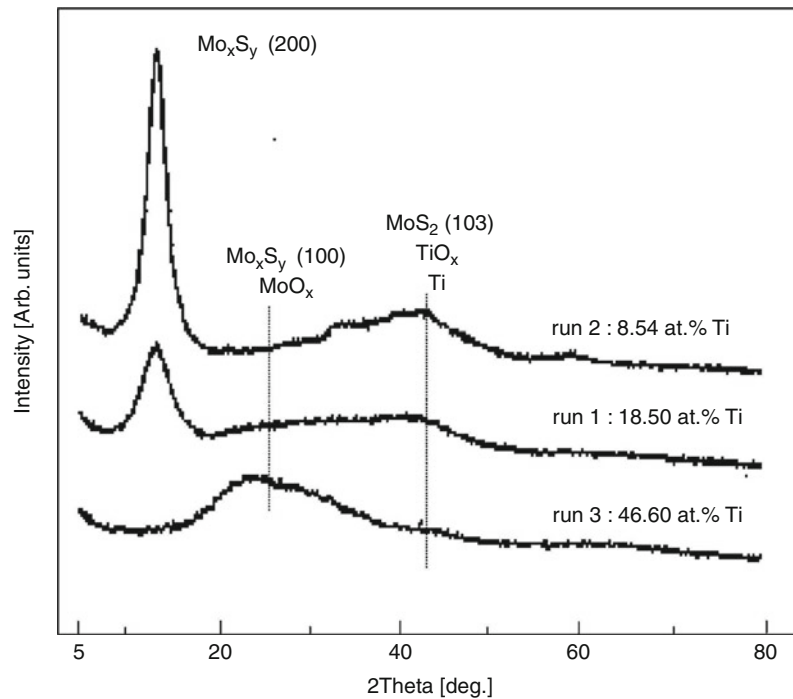
B-based coating with superimposed MoS₂, TiC, and TiB₂ line spectra. The MoS₂ phase is obviously nanocrystalline and shows a more ordered (002) structure, which is expected to be a basal plane parallel to the substrate. Due to the low atomic numbers of B and C, the Ti containing phases are expected to dominate in the composite

structure. The preferred peaks are (100) for TiB₂ and (111) for TiC, which are shifted due to close diffraction angles.

Microhardness

The microhardness of MoS₂-Ti coatings coated by pulsed-dc changed with Ti and S/Mo ratios. The highest microhardness value was obtained with the film having a thickness of 1.7 μm, S/Mo ratio of 1.08, and 8.54 at.% Ti. The high microhardness of MoS₂-Ti coating (4.5 GPa) is attributed to its denser (see Fig. 3) structure and oriented basal plane (0 0 2) (see Fig. 6). It can be said that MoS₂-Ti films having low Ti content (less than about 15%) and S/Mo ratio higher than 1 exhibit the pronounced basal orientation and the highest microhardness (Arslan et al. 2005; Bulbul et al. 2007).

MoS₂-Nb coated AISI 52100 tool steel substrates were tested for microhardness values. The highest microhardness determined is ≈5.5 GPa. There is no previously published literature for the MoS₂-Nb self-lubricating films to compare this result to. As shown in Fig. 3, it was observed that MoS₂-Nb coatings have a very dense microstructure. The highest microhardness value was obtained with the film having a thickness of 2.5 μm, S/Mo ratio of 1.66, and 7.82 at.% Nb. The microhardness test results have shown that Nb addition increased the film hardness (Efeoglu et al. 2008; Arslan et al. 2008). These results have indicated that Ti- and Nb-added MoS₂ solid lubricant thin films showed very high microhardness.



Sputtering MoS₂-based Coatings, Fig. 6 XRD multiplot spectra of MoS₂-Ti coatings (Taken reference Efeoglu and Bulbul 2005)

The measured microhardness of TiN/TiAlN-MoS₂/MoS₂-Ti and TiB₂C:MoS₂ multilayered coatings are about ≈ 4.5 GPa and ≈ 7.8 GPa, respectively (Efeoglu 2005; 2007;).

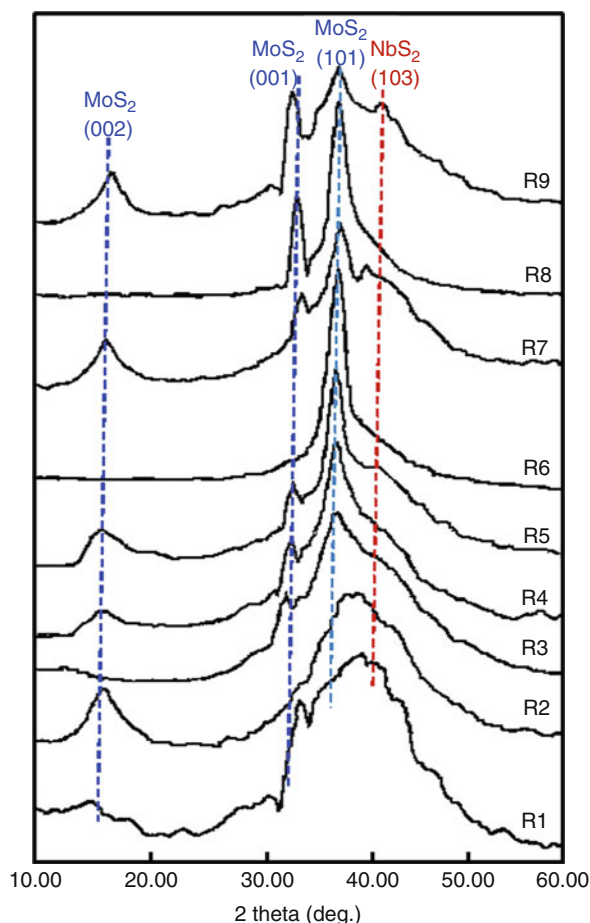
Adhesion and Fatigue Properties

In order to determine the number of cycles until failure occur (fatigue properties), a multi-pass scratch test (Revetester by CSM Instruments) was used under the sub-critical load and the critical load value on the MoS₂-Nb coated sample. A number of scratches (unidirectional) were performed at a given load without removing the specimen (Arslan et al. 2008). The initial critical load, L_{C1} , is recorded when the film is pulled off from the substrate with adhesive failure of the flaking or with cohesive failures from the coating itself. While thinning of the coating is under progress, there is a sudden release of the energy and a change of the shear stress because the diamond stylus reaches the substrate. This point corresponds to the first contact of the diamond stylus with the substrate. At this contact point, the critical load characteristic of total failure of the coating is called the final L_{C2} (see Fig. 10).

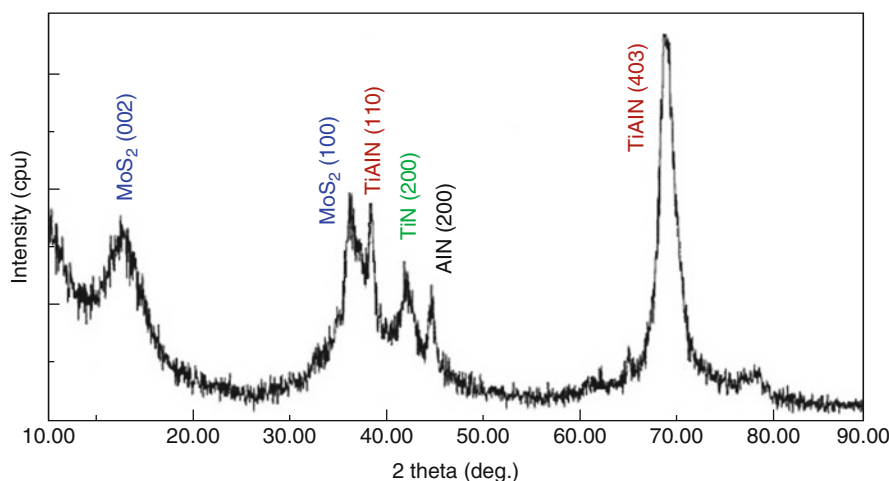
MoS₂-Nb: The measured normal load (F_n) and corresponding (F_t) values with acoustic emission (AE)

were plotted. Figure 11 is a typical example. The critical load to failure, which was determined using a combination of acoustic emission and the increase in the friction coefficient, was shown to be 15 N as first L_{C1} . In the scratch test investigation, pre-critical load flaking of the coating at the edge of the scratch channel was observed with MoS₂-Nb, while no such flaking was found when the applied load increased to 15 N. After thinning of the coating, the diamond stylus reached to the substrate at 95 N load, which is called the final critical load (L_{C2}). It can also be seen that the friction force increases irregularly after reaching final value of L_{C2} (~ 95 N) (see Fig. 11). The stylus pressed the film into contact with the substrate and caused the delaminating of the film from both edges of the scratch track as recovery spallation. The multi-pass scratch test was performed at 1,000 cycles across the same scratch line at 5, 8 (under L_{C1}) and 15 N force (critical load value).

After many unidirectional passes across the MoS₂-Nb film surface, the differences between the friction coefficients as function of the number of the cycles to failure are seen in Fig. 12. It is not expected that the major damage mechanism for unidirectional sliding is side adhesive flaking, but the crack density does not seem to increase as a function of cycle number. For each coated surface, 50, 100, 500, and 1,000 passes were completed, and



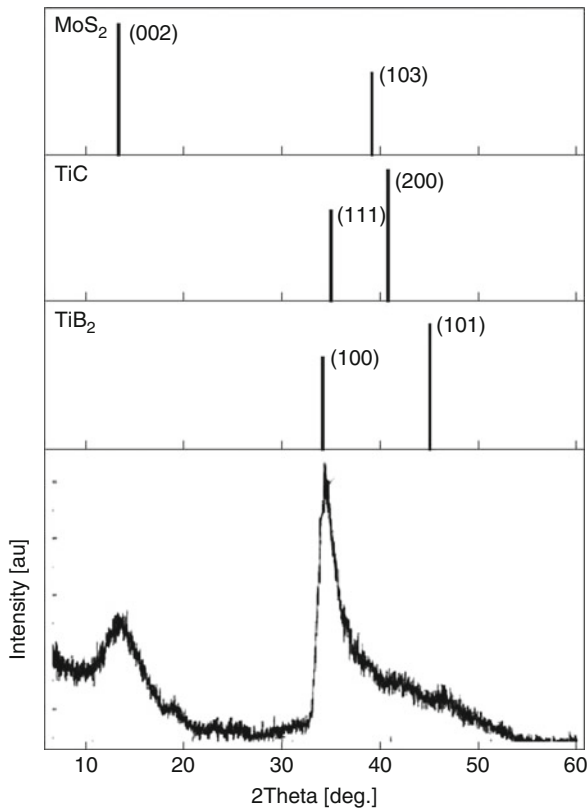
Sputtering MoS₂-based Coatings, Fig. 7 XRD analysis from MoS₂-Nb coating (Taken reference Efeoglu et al. 2009a)



Sputtering MoS₂-based Coatings, Fig. 8 XRD analysis from TiN/TiAlN-MoS₂/MoS₂-Ti multilayer coating (Taken reference Efeoglu 2007)

subsequent scratch tracks are shown in Fig. 13. There are typical fatigue failure appearances, where the flaking particles have taken the shape of flat plates after 500 cycles at 15 N. Perhaps the most significant finding in the test is that when the multi-scratch passes reached to 1,000 cycles, micro-scale fatigue failures disappeared. It is assumed that the produced micro-scale solid lubricant debris accumulated in that failure regions to create a new self-lubricating film (Arslan et al. 2008).

TiN:TiAlN-MoS₂:MoS₂-Ti: Figure 14 shows the result of a typical scratch test from the multilayered coating system. The critical load for the multilayer-composite system varied between 80 and 90 N. The test results confirmed the excellent adhesion of the multilayer-composite coating to the D2 tool steel substrate (Efeoglu 2007). As far as the coating failures are concerned, the coating showed no damage in the scratch sides and into the scratch path. Stresses in the coating at scratch sides are related to the compressive stress fields. As seen in Figs. 14 and 15, there are no adhesive and/or cohesive failures, while the coating is thinning under the indenter tip. Figure 15 shows a very clear progression of the thinning of the multilayer with no observable microcracks or defects at the interface between MoS₂-Ti and TiAlN-MoS₂ layers. As the thinning process reaches to TiN and then to the substrate, appear. It must be noted here that the multilayered structure dispersed the accumulated stress in its structure by plastic deformation and it is believed that the stress evolution was interrupted at the layer boundaries, which positively affected the adhesion property of the coating (Lee et al. 2003).



Sputtering MoS₂-based Coatings, Fig. 9 X-ray diffraction patterns of MoS₂:TiB₂:Ti based coating (Taken reference: Efeoglu 2005)

TiB₂:TiBN-MoS₂:MoS₂-Ti: A new aspect of TiB₂:TiBN-MoS₂:MoS₂-Ti based graded multilayer coatings is the investigation of interfacial adhesion behavior using the scratch test system. The multilayered-composite solid lubricant coatings were deposited from separate Ti, TiB₂, and MoS₂ targets. Figure 16 shows the result of a typical scratch test from the multilayered coating system (Efeoglu et al. 2009b). The thinning of the coatings by plastic deformation occurred from first interface through third interface. Layer spalling-like adhesive failure on both sides of the scratch tracks was seen at the MoS₂-Ti (top layer). The coatings showed no interfacial spalling/buckling failures at the first L_c points. The critical loads occurred as function of the film thinning process. The scratch test results indicated that, while the substrate bias voltage (−50 V), working pressure (0.26 Pa), and duty time (2 μs) were kept constant, the interfacial adhesion values changed as function of the nitrogen flow rate and the pulses applied to the substrates. All interfacial critical

loads decreased with increasing nitrogen flow rate and the pulse applied to the substrates.

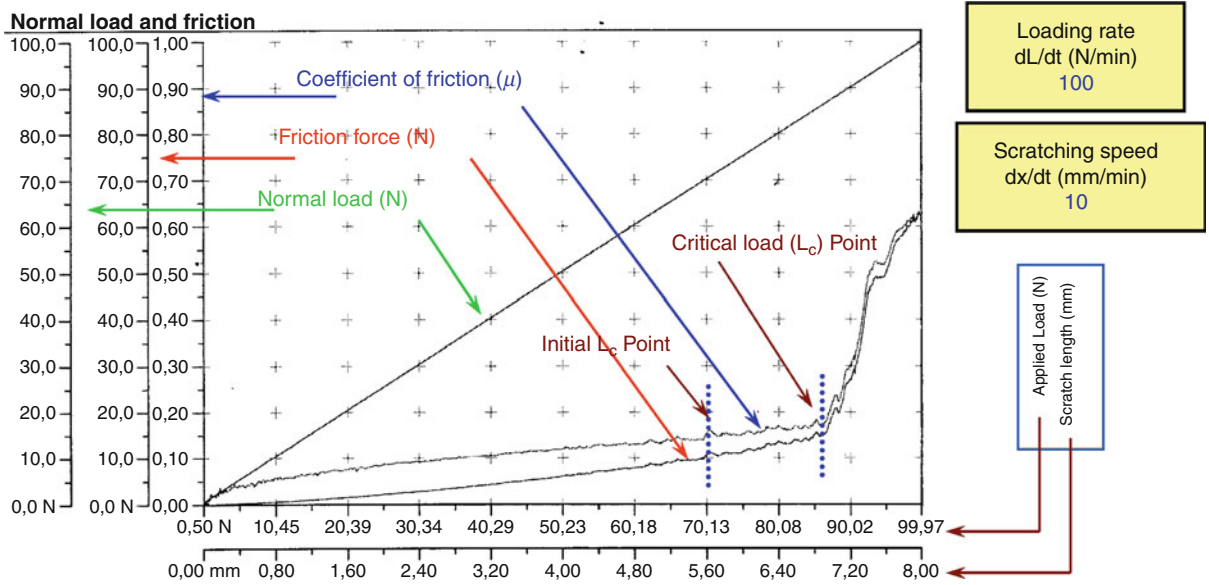
Tribological Properties

The pin-on-disc was used to investigate friction and wear behavior of the MoS₂-based composite films. Tribo-tests were conducted at different conditions, and in dry nitrogen. Wear tracks were characterized with a profilometer. Wear volume was calculated using the cross-sectional area from the wear track profile, and thus the wear rate was attained using the equation $K = V/ws$, where K is the value of the wear rate, V is the worn volume, w is the normal load, and s is the distance moved.

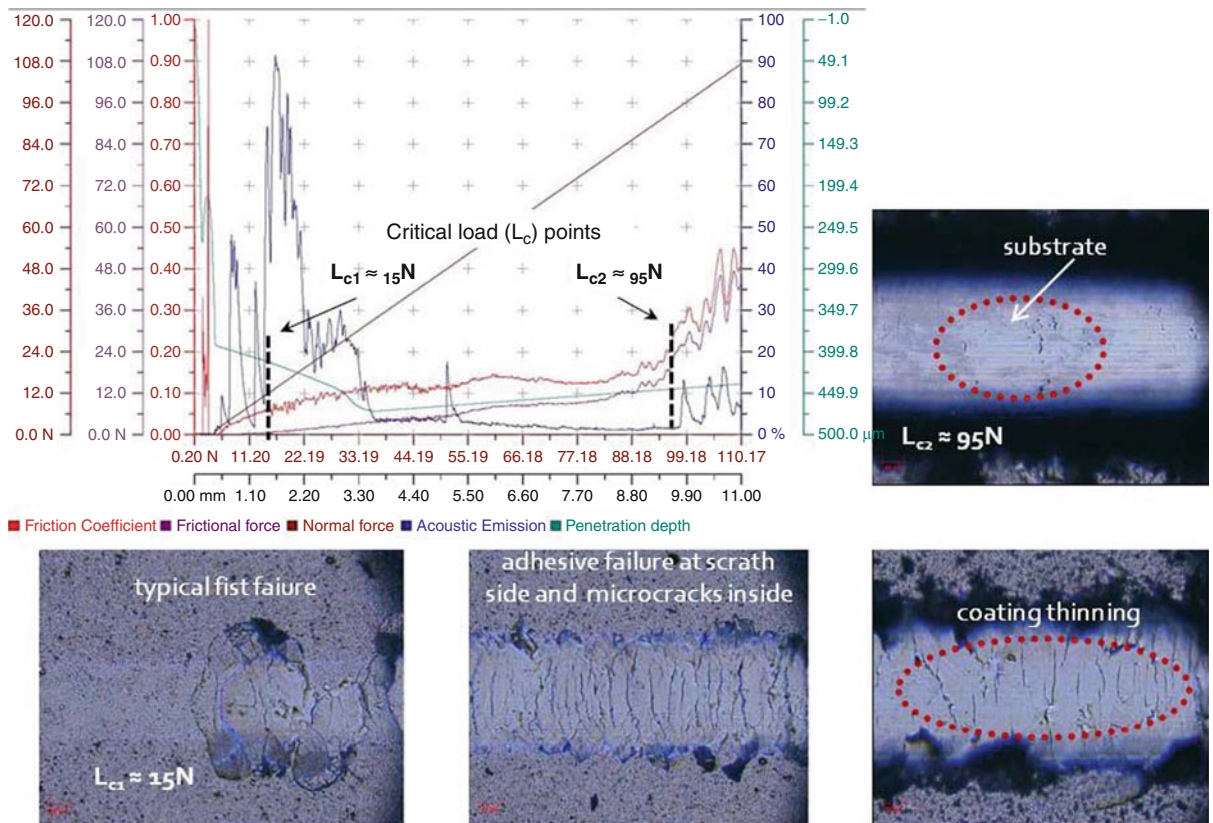
It can be said that the (002) basal plane parallel to the substrate, which has low energy and is inert chemically, which is preferred for tribological applications, formed partially. Many authors reported that a well-adhered transfer film occurs between the surfaces sliding on each other, and this transfer film results in a low friction coefficient (Fleischauer 1987). One example of the formation of the transfer film layer on a counter-surface is shown in Fig. 17 (Arslan et al. 2004).

MoS₂-Ti: The sliding wear life for the films deposited having low Ti content (8.54 at.%) in the investigations were determined in terms of the number of passes at which CoF was between 0.055–0.070 and 0.022–0.035 in humid air and in dry nitrogen, respectively (Arslan et al. 2004; Efeoglu and Bulbul 2005). The effects of deposition parameters on stoichiometry, the structural change occurring with Ti incorporation, and the influence of structural change on CoF for MoS₂-Ti composite film have been examined and some conclusions were drawn: (a) While unstable friction coefficients were obtained in the films in which MoO_x formed and thus with high (100) peaks, lower friction values were achieved for the films having pronounced (002) basal plane orientation. (b) Reactive (100) plane orientation sensitive to environmental attacks was tolerated with Ti incorporation. MoS₂ matrix structure variation for MoS₂-Ti coatings deposited at different deposition parameters has been demonstrated. (c) By contributing to MoO_x formation, reactive pronounced (100) reflections caused an increase of wear rate. However, this undesirable effect of reactive (100) plane orientation has been eliminated with Ti incorporation, so that it provided lower wear rates by the denser and stronger structure obtained.

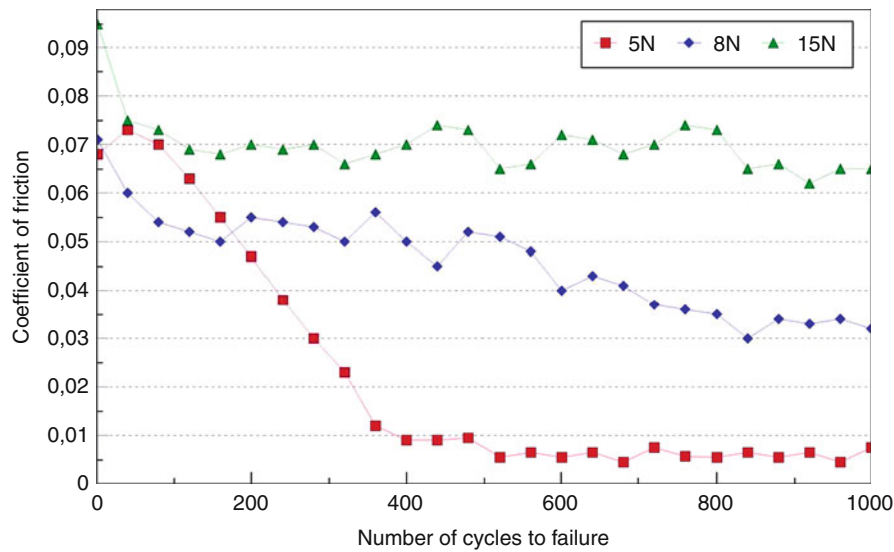
The results from the friction and the wear results indicate that CoF and the wear life of MoS₂-Ti films decreased as a function of the environmental conditions. Figure 18 presents typical wear life traces obtained in humid air and in dry nitrogen for pulsed-dc magnetron



Sputtering MoS₂-based Coatings, Fig. 10 Typical scratch adhesion test output from a thin film coated surface (Unpublished figure)



Sputtering MoS₂-based Coatings, Fig. 11 Typical scratch adhesion test output from MoS₂-Nb coated surface (Unpublished figure)



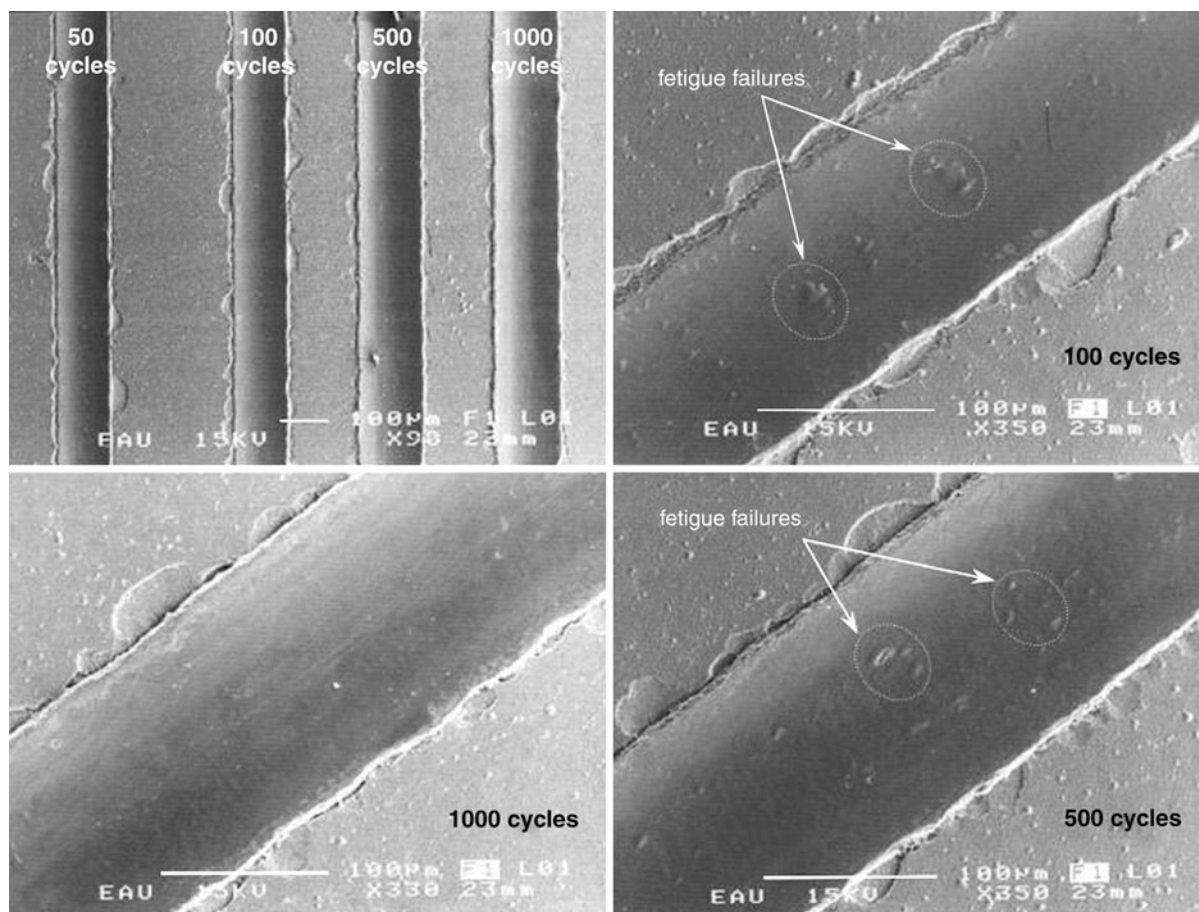
Sputtering MoS₂-based Coatings, Fig. 12 Variation of friction coefficient with number of cycles and applied loads for MoS₂-Nb (Taken reference Arslan et al. 2008)

sputtered MoS₂-Ti. The sliding wear life varied with the environment and the preferred crystallographic orientation. The highest wear life of the MoS₂-Ti films (0.03E-6 mm³/Nm) were over 250,000 passes in dry nitrogen but only 142,500 passes in humid air (2.11E-6 mm³/Nm) where the film was (0 0 2) basal plane oriented.

MoS₂-Nb: CoF for MoS₂-Nb films that are measured in air, dry nitrogen, distilled water, and oil conditions are shown in Fig. 19. Both the highest friction coefficient and the wear rate were attained on MoS₂-Nb coating tested in dry nitrogen. Also, the lowest friction coefficient and wear rate were attained in the synthetic oil (Efeoglu et al. 2008).

The tribological properties of MoS₂ degrade in the presence of humidity. Therefore, it was generally expected that CoF in dry nitrogen should be smaller than the one in the humid air condition. These experimental results are completely different from the well known MoS₂-Ti type solid lubricant films. In Nb-doped MoS₂, growth of Nb_xS_y phases is possible (see Fig. 7). Niobium sulfides are part of the family of transition metal dichalcogenide material such as MoS₂. NbS₂ has been shown to be an effective humidity sensor (Divigalpitiya et al. 1990). Oxygen affinity of NbS₂ seems to have a beneficial effect on the MoS₂-Nb self-lubricated film. Oxygen trapping is faster in NbS₂ than in MoS₂-Nb phases. Therefore, it is also possible that the oxidation reaction in humid air condition takes place in Nb_xS_y rather than the MoS₂-Nb (002) phase.

Figure 20 illustrates the relationship between CoF and lap at room temperature and different temperatures for MoS₂-Nb composite coatings (Arslan et al. 2010). It has been observed that CoF significantly changed depending on the temperature. It was also determined that, while a very stable CoF was measured at about $\mu \approx 0.072$ from the wear tests performed at room temperature, an unstable CoF was observed from the tests carried out at 300°C. This unstable CoF changed between 0.035 and 0.061. The reason for this unstable behavior of the CoF was associated with the formation of MoO_x by rapid oxidation of the MoS₂ (001) plane, which is sensitive to oxidation with an increase in temperature. Furthermore, the formation of the Nb_{1-x}S phase, which has a lower solid lubricant property compared with that of MoS₂, caused unstable wear. It was observed from the high-temperature tribological test performed at 500°C that the coating showed a rather insufficient tribological behavior. The CoF at this temperature, which was measured as 0.19 at the beginning, depending on the Hertzian contact, gradually increased up to about 1,000 laps. In this interval, it was observed that the CoF of the coating increased from 0.067 to 0.14. The mean CoF at this temperature was $\mu \approx 0.014$. This low CoF might be attributed to the nonexistence of a Nb_{1-x}S phase at this temperature. From the observation of in situ XRD patterns (see Fig. 21, although Nb_{1-x}S phases formed intensively at 500°C), sulfur did not diffuse into the coating at 100°C, and thus the Nb_{1-x}S phase did not form. The fact that the Nb_{1-x}S phase did not form at

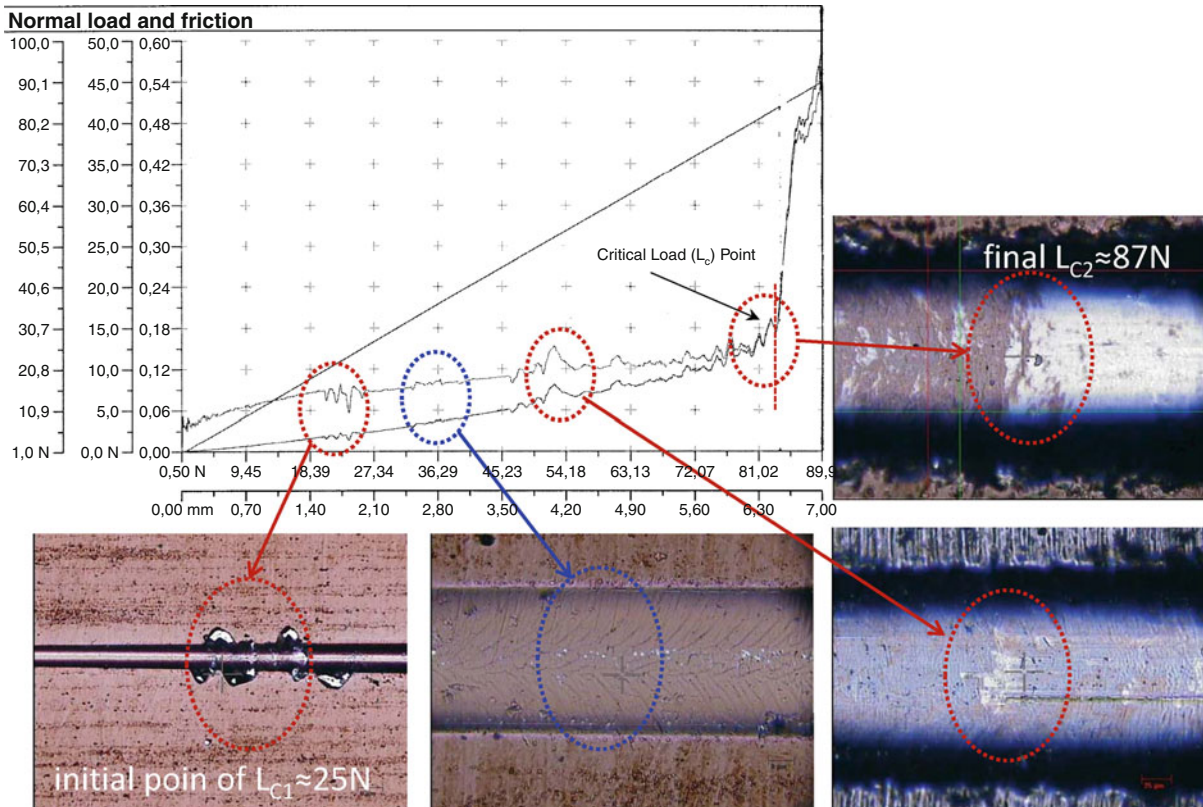


Sputtering MoS₂-based Coatings, Fig. 13 The coating condition after multi-cycles under 15 N on MoS₂-Nb coating (Taken reference Arslan et al. 2008)

this temperature resulted in a stable and low friction coefficient.

TiN:TiAlN-MoS₂:MoS₂-Ti: The sliding wear life for the coatings in this investigation was determined in terms of the number of passes at which the initial coefficient of friction was between 0.06 and 0.07 at the end of the 1,100 s sliding time. As the sliding time increases to 3,600 s, the CoF fell to a low, constant value (μ :0.06) maintained until the end of the test. As seen Fig. 22, EDS data were taken from three positions in the wear track where the worn surface from the multilayered-composite indicates three worn zones. A qualitative analysis by EDS confirmed the presence of Ti (at.% 22), Al (at.% 10), N (at.% 11), Mo (at.% 25), and S (at.% 32) within the multilayered-composite coating. All worn debris piled up at the both side of the wear track with no abrasive wear effect. This is the reason

why CoF remained very stable during the 1 h sliding test. It is well known that the main reason for low CoF seems to be the presence of a composite structure of TiAlN-MoS₂ and MoS₂-Ti top layer. Otherwise, it has been noted that one of the reasons for the low CoF obtained by tribotest is the oxidation resistance of the TiAl-based phases in the coating structure. TiAl₃ is known as an alumina former with reasonably good oxidation resistance. TiAl₃ is generally considered as a coating material for oxidation protection of the Ti₃Al- and TiAl-based materials. The coating produced as multilayered-composite film showed quite high wear resistance owing to the middle composite layer (TiAlN-MoS₂) and adhesion strength. The calculated wear rate ($0.35 \times 10^{-6} \text{ mm}^3/\text{Nm}$) was taken after 9,000 revolutions of the tribotest. This value is lower than previous works (Arslan et al. 2005; Efeoglu and Bulbul 2005).



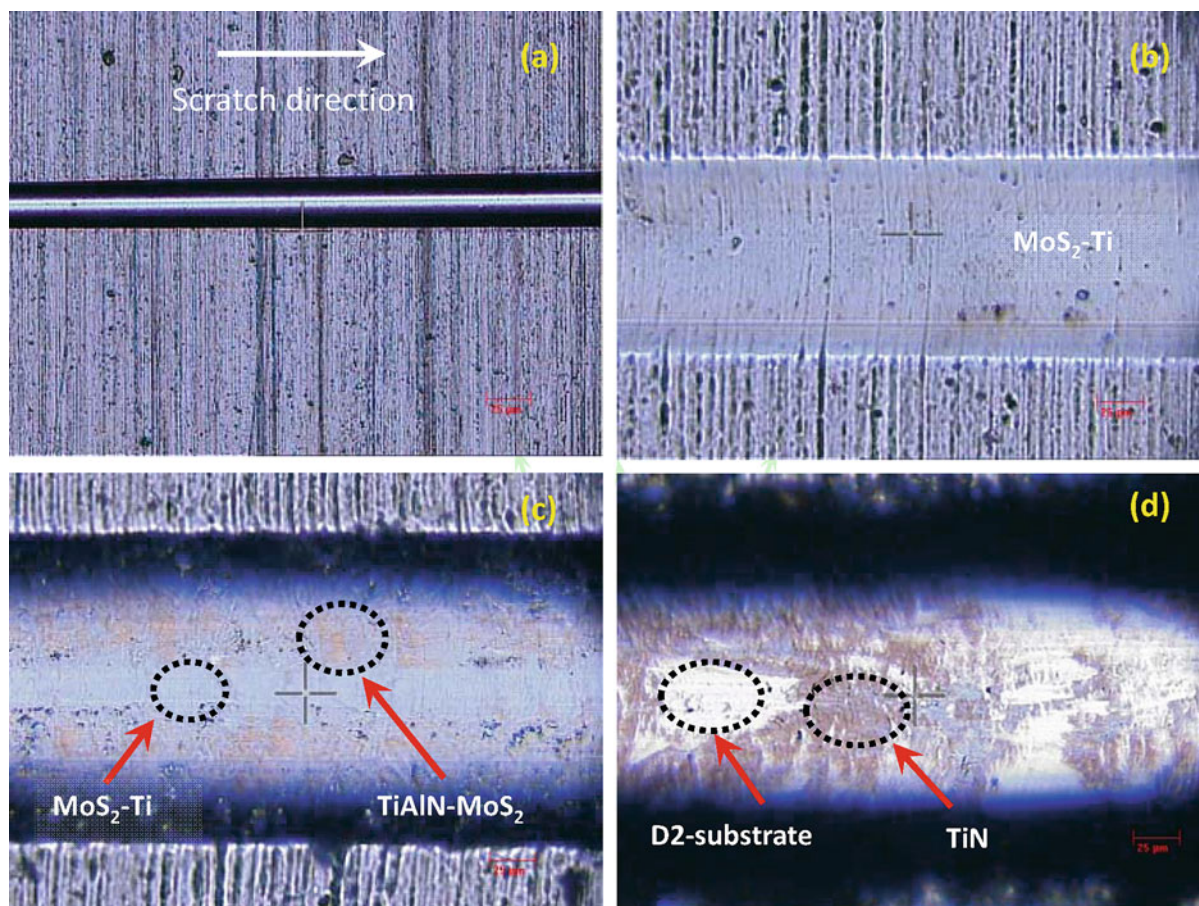
Sputtering MoS₂-based Coatings, Fig. 14 Typical scratch adhesion test output from TiN-TiAlN/MoS₂-Ti coated surface (Unpublished figure)

TiB₂:C:MoS₂: With the Revetest instrument, the friction and wear performance of a coating has been investigated by means of a multi-pass scratch test. This procedure consists of sequential constant load sliding scratches over the same track with a low load. Bidirectional scratch testing was performed using a fully automatic CSM-Revetester. Illustrations of the scratch track morphology after 500 cycles under different loads are given in Fig. 23 (Efeoglu 2005). Debris is clearly visible at the edges of the track, in spite of the low load applied. Compacted and loose debris become visible along the sides of the tracks from 40 to 60 N. It can be seen that the amount of compacted and loose debris deposited along the sides of the wear tracks increased with the applied load. Initially, no damage is found under 10 and 20 N applied loads. The wear tracks on the co-sputtered Mo:S:C:Ti:B-based coatings reveal that the sliding action usually generated a smooth wear surface. However, SEM

images show localized micro-buckling and delamination of Mo:S:C:Ti:B-based coatings after 40 N × 500 cycles. It is assumed that while adhesive failures, as small debris, coming up due to low cycle fatigue at the worn surface, were easily trapped in the contact area as a third body until the patches of the compacted debris (up to 5,000 cycles) could play the lubricating role, that resulted in the sudden decrease of friction coefficient ($\mu \approx 0.052$) up to 7,000 cycles. The following rapid increase of friction coefficient could be explained by adhesively delaminated worn Mo:S:C:Ti:B-based coating. No signs of damage were observed even after 500 and 1,000 cycles.

Results and Discussion

Different compositions of MoS₂-based single/multilayered composite coatings deposited by changing deposition parameters, and the effect of deposition parameters on stoichiometry, the structural change occurring with Ti and

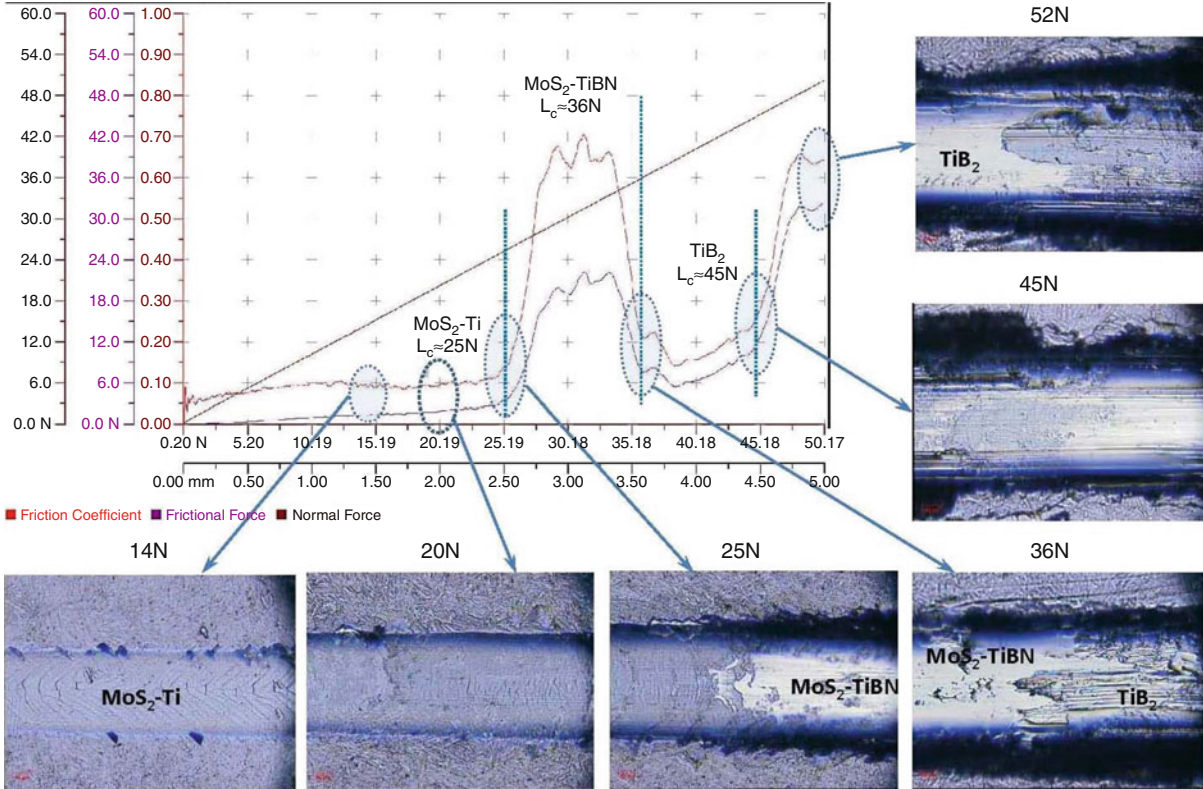


Sputtering MoS₂-based Coatings, Fig. 15 Typical scratch adhesion test output from TiAlN/MoS₂-Ti (Taken reference: Efeoglu 2007)

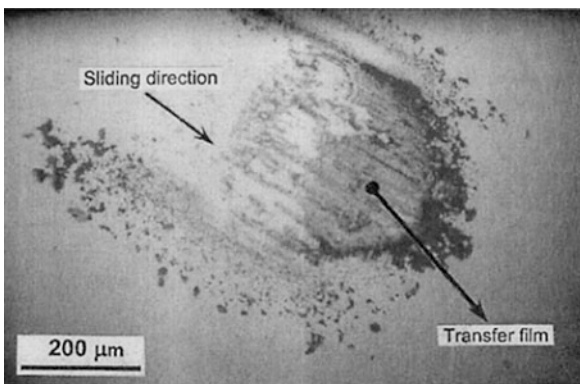
Nb incorporation, and the influence of structural change on mechanical and tribological properties were examined. The following conclusions have been drawn:

- Remarkable results were achieved in terms of coating quality with application of pulsed-dc power to the substrates. The coatings show very dense microstructure with pronounced basal plane orientation. The coating showed the lowest friction coefficient (0.022) in dry nitrogen due to the strong oriented basal plane (2 0 0) parallel to the surface. (0 0 2) plane intensity changed depending on working pressure and Ti content. This plane was well aligned parallel to the substrate at the working pressures of 0.26 and 0.33 Pa. The lowest Ti content (about 8.54%) and the strongest (0 0 2) basal plane orientation were observed with the lowest Ti target current (0.5 A) and the highest MoS₂ target current (0.9 A), although the highest Ti content (about 46.63%) and no basal plane orientation were observed with the highest Ti target current (1.5 A) and the lowest MoS₂ target current (0.3 A). With increasing Ti concentration in the Mo × Sy matrix, the structure becomes amorphous.
- The high microhardness of MoS₂-Ti (4.5 GPa) is attributed to its denser structure and oriented basal plane (0 0 2). There is very clear evidence that Nb addition increased the film hardness more than Ti addition. The measured highest surface hardness increased to 5.5 GPa.
- It was noted that the highest N_S/N_{Mo} atomic ratio (1.59) was obtained at the working pressure of 0.40 Pa and −30 V bias voltage, and the lowest was at 0.26 Pa and −70 V bias voltage.

Normal load and friction

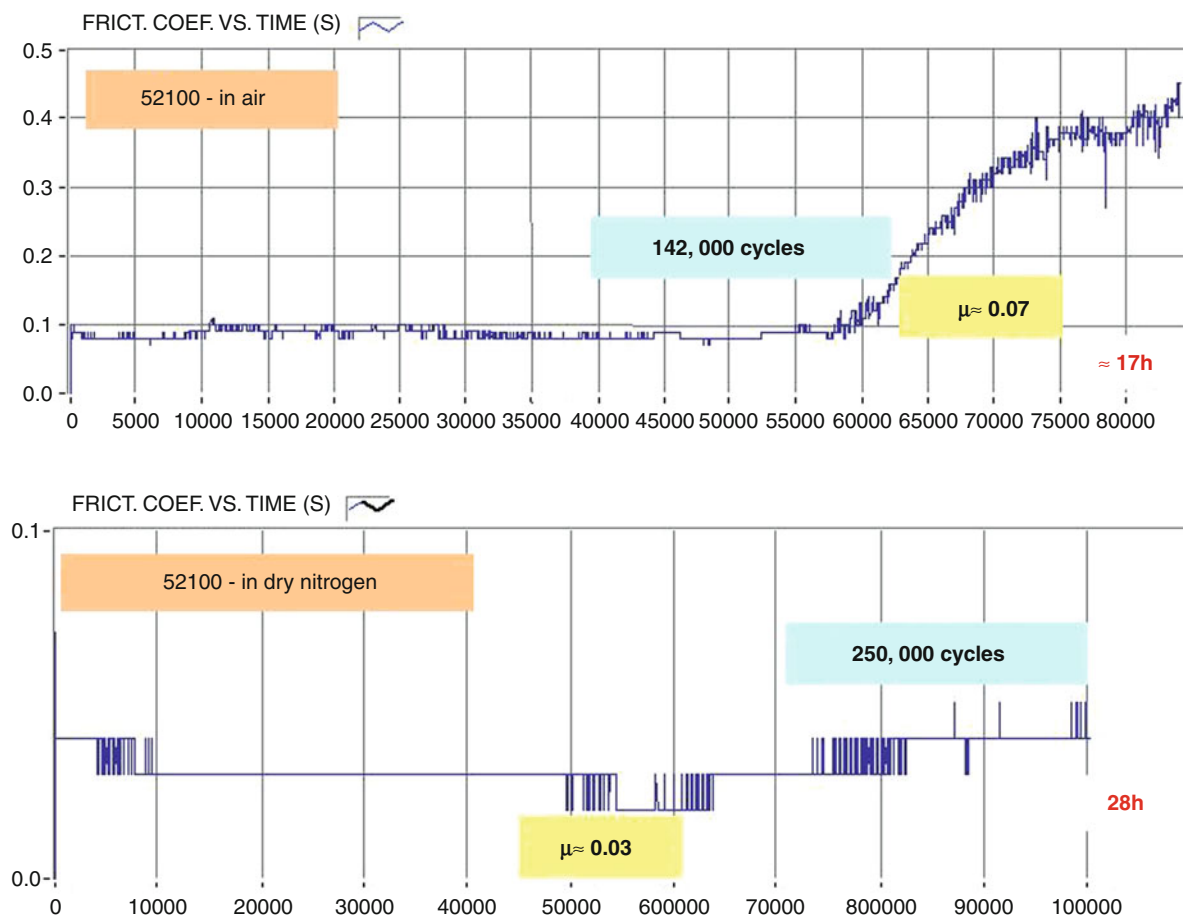


Sputtering MoS₂-based Coatings, Fig. 16 Typical critical load points taken when scratching TiB₂:TiBN-MoS₂:MoS₂-Ti coating (Taken reference Efeoglu et al. 2009b)



Sputtering MoS₂-based Coatings, Fig. 17 Transfer film layer formed on counter surface during sliding (Taken reference Arslan et al. 2004)

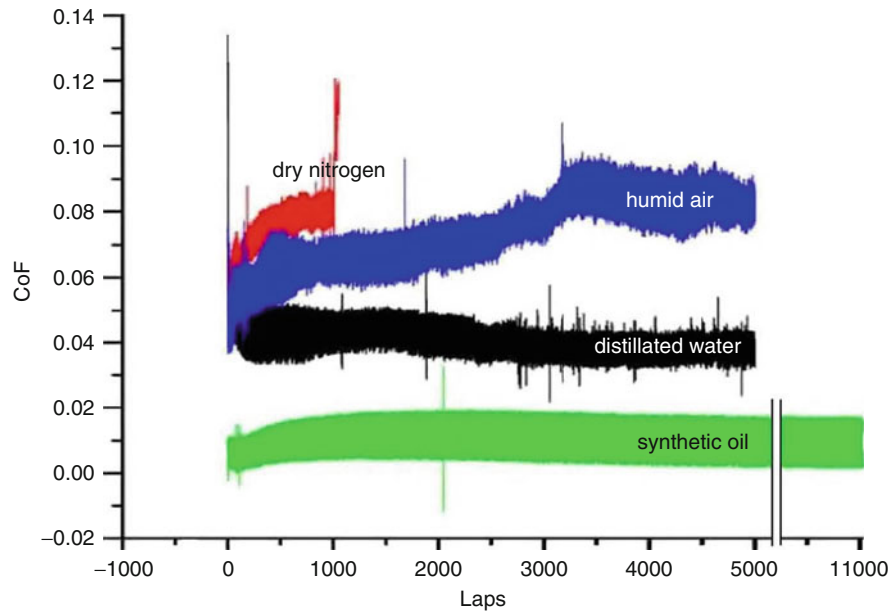
- In humid air, an oxidative environment, the films had higher coefficient of friction (0.07), and lower wear life (142,500 passes with $2.11 \times 10^{-6} \text{ mm}^3/\text{Nm}$ wear rate) than in dry nitrogen, where the wear life was over 250,000 passes with the lowest wear rate ($3 \times 10^{-8} \text{ mm}^3/\text{Nm}$).
- MoS₂-Nb has the lowest CoFs, wear rate, and showed no ball wear. CoFs have been observed as 0.075, 0.085, 0.04, and 0.01 in humid air, dry nitrogen, distilled water, and synthetic oil conditions, respectively.
- Pulsed-dc sputtered MoS₂-Nb films exhibited two types of crystallite orientation and characteristic very dense structure corresponding to oriented growth in the low-shear basal plane of (002) and humidity-sensitive NbS₂ phase. Furthermore,



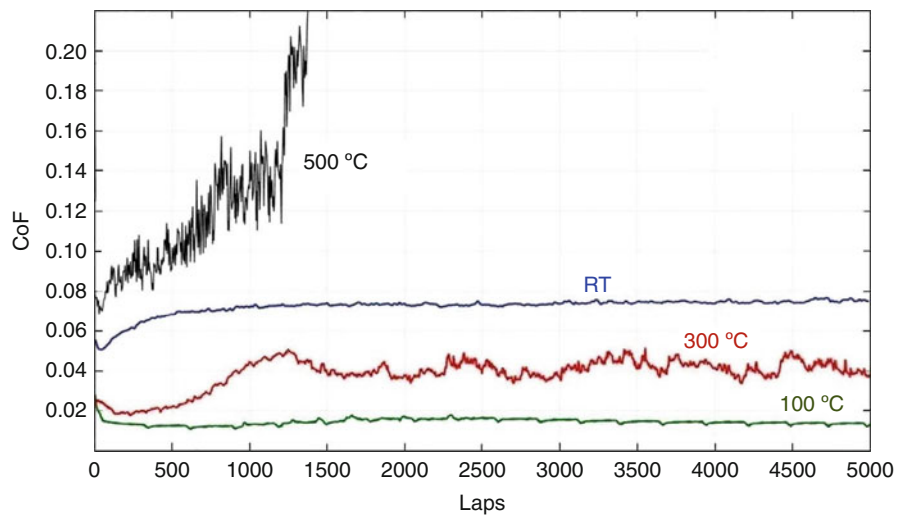
Sputtering MoS₂-based Coatings, Fig. 18 Friction coefficient as function of the sliding time on MoS₂-Ti coated 52100 steel substrate (a) in air condition and (b) in dry nitrogen (Taken reference Efeoglu and Bulbul 2005)

tribological performance of the MoS₂-Nb self-lubricating films was much better in humid air than in dry nitrogen.

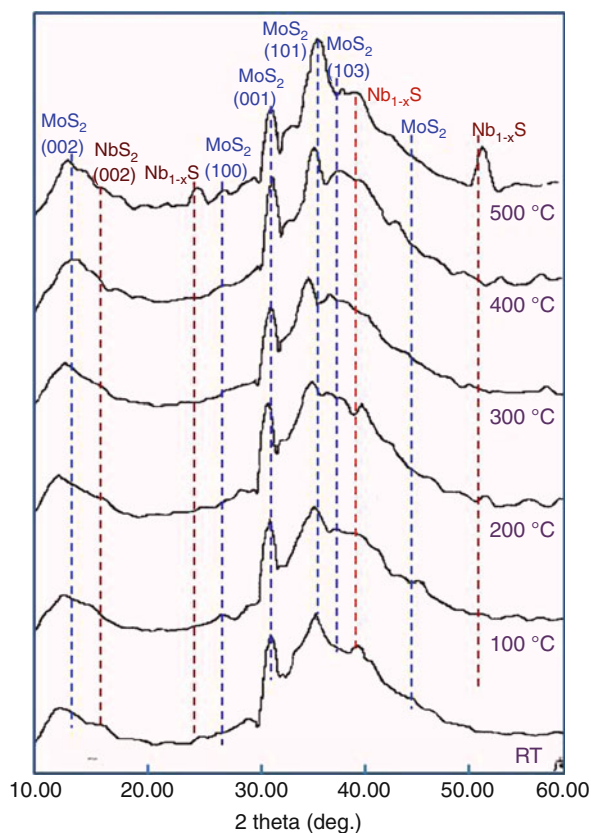
- CoF was measured at about $\mu \approx 0.014$ at 100°C. At the temperatures of 300°C and 500°C, rapid oxidation of the coatings with the temperature and formation of an Nb_{1-x}S phase with a high-friction coefficient compared to MoS₂ resulted in the increase of CoF. In addition, while the lowest wear rate was determined from the wear test at 100°C, it was observed that this rate increased for the other temperatures.
- In the scratch test investigation, pre-critical load flaking of the coating at the edge of the scratch channel was observed with MoS₂-Nb, while no such flaking was found when the applied load increased to 15 N. Acoustic emission data corresponds to a load of 95 N.
- The CoF from the multi-scratch for MoS₂-Nb started at a very high value of around 0.067, 0.073, and 0.093 for 5, 8, and 15 N loads, respectively, and then dropped to 0.006, 0.035, and 0.065 in the applied loads of 5, 8, and 15 N, respectively, at the end of 1,000 cycles. There are typical fatigue failure appearances, where the particles have taken the shape of flat plates after 500 cycles at 15 N. Perhaps the most significant finding in the test is that when the multi-scratch passes reached to 1,000 cycles, micro-scale fatigue failures disappeared.



Sputtering MoS₂-based Coatings, Fig. 19 Variation of friction coefficients in different tribo-test conditions as function of the laps (Taken reference Efeoglu et al. 2008)



Sputtering MoS₂-based Coatings, Fig. 20 Changes in the friction coefficients at high temperatures of the MoS₂-Nb composite solid lubricant coatings as a function of lap (Taken reference Arslan et al. 2010)



Sputtering MoS₂-based Coatings, Fig. 21 In situ XRD patterns at different temperatures of the MoS₂-Nb composites solid lubricant coatings (Taken reference Efeoglu et al. 2009a)

It is assumed that the produced micro-scale solid-lubricant debris accumulated in that failure regions to create a new self-lubricating film.

- The critical load for the multilayer-composite system (TiN:TiAlN-MoS₂:MoS₂-Ti) varied between 80 and 90 N. The test results confirmed the excellent adhesion of the multilayer-composite coating.
- Interfacial adhesion behavior from TiB₂:TiBN-MoS₂:MoS₂-Ti type multilayered films scratch test results showed that the thinning of the coatings by plastic deformation occurred from the first interface through third interface. Layer spalling-like adhesive failure on both sides of the scratch tracks was seen at the MoS₂-Ti (top layer). The coatings showed no interfacial spalling/buckling failures at the first L_c points.

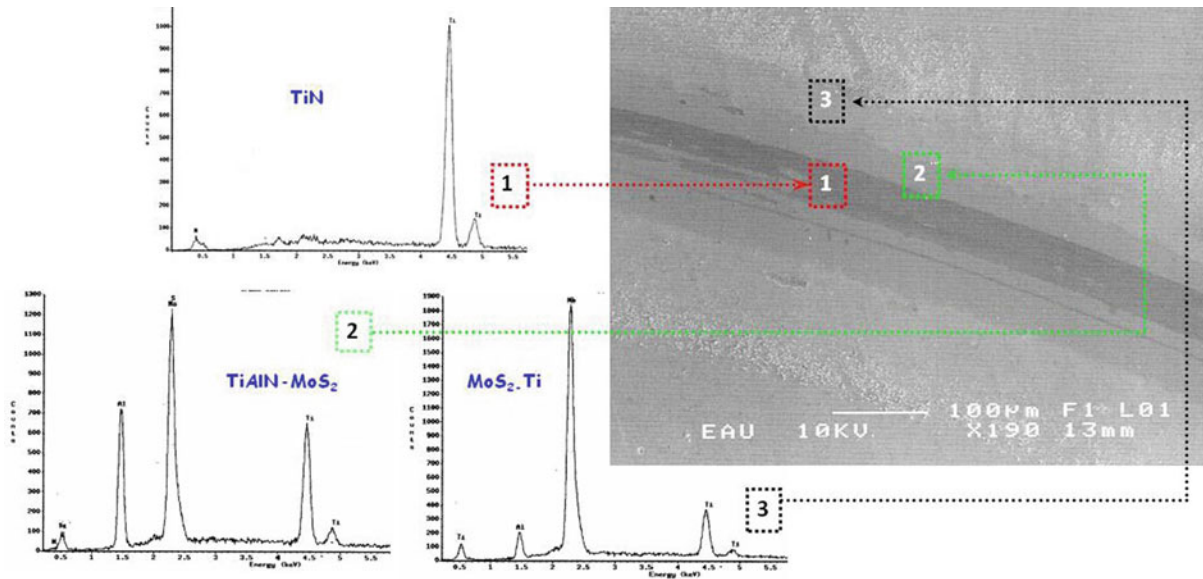
- Co-sputtered TiB₂:C:MoS₂ based coatings with dense and noncolumnar microstructure showed a more ordered (002) structure. The measured CoF was noted as low as 0.042. These types of coatings show a high load-carrying capacity. When the maximum stresses in the contact region exceed some critical value, severe plastic deformation due to low cycle fatigue becomes the dominant spalling plastic deformation after 4,000 cycles. It is important to note that the patches of compacted debris at the ends of the tracks are directly under the turnaround points at both ends. It was noticed that transfer layers were not developed on the pin surfaces after 8,000 cycles at all applied loads. Moreover, it was not possible to detect any measurable worn scar on the pin after 8,000 cycles.

Key Applications

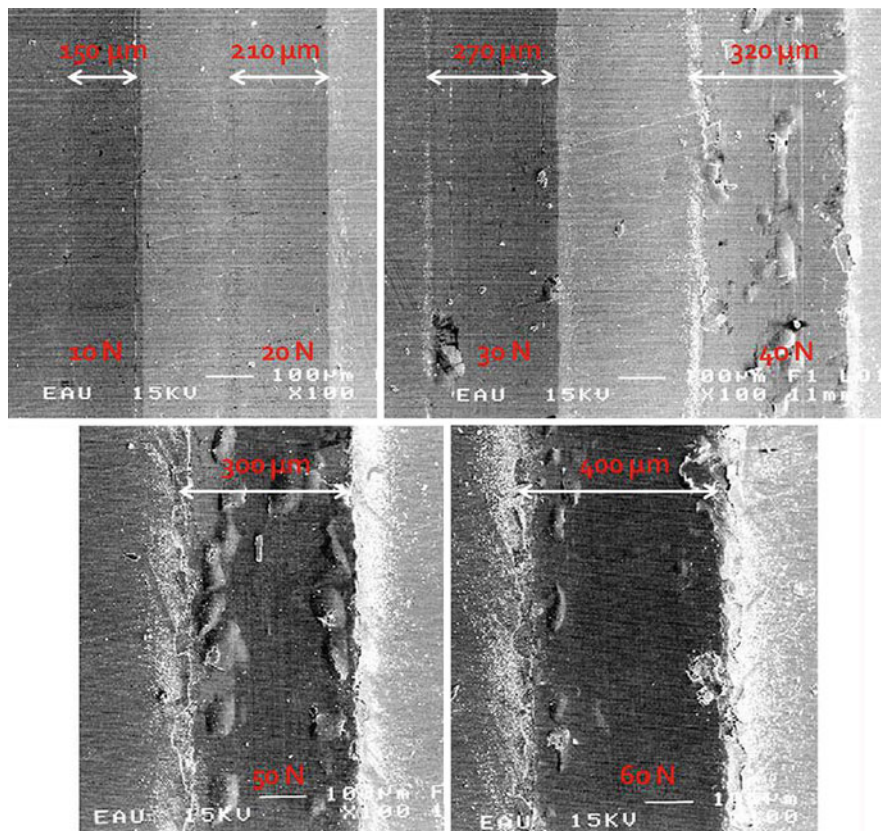
Historical factors and industrial evolution have influenced the lubrication techniques applied to most machines. In particular, dry solid lubrications of sliding surfaces by use of solid thin films/coatings have become a common method in industrial applications. As a solid lubricant, MoS₂ is well known for its lamellar structure with weak van der Waals bonding between crystal planes. It has been recognized that friction and wear performances of sputtered MoS₂ and its composite form depend on the deposition conditions and the operating environment. Due to their good performance in space environment, MoS₂ films have been used for space applications. In addition, MoS₂ solid lubrication films will have many non-space applications where dry lubrication is required, including X-ray rotation anodes, cryogenic coolers, air-bearing environments, and cutting tools, and as photoactive materials in solar energy conversion.

Future aeronautical and space missions will use surface engineering technologies applied to advanced aircraft and spacecraft based on very low and high temperature, high vacuum, nuclear/cosmic radiation, and aggressive environments. Some new applications such as microelectromechanical systems (MEMS), gears, and bearings for long-term service in space and atmospheric mechanisms have prompted renewed interest in solid lubricant thin films.

Today, the ideal solid lubricant coatings would combine the hardness of diamond, the toughness of high-speed steel, the chemical inertness of alumina, and excellent adhesion to the substrate, with the friction coefficient as low as 0.001.



Sputtering MoS₂-based Coatings, Fig. 22 EDS analysis on the wear scar from the multilayered coated film (Taken reference Efeoglu 2007)



Sputtering MoS₂-based Coatings, Fig. 23 An SEM image of wear tracks of MoS₂:C:Ti:B based coatings obtained at various applied loads and 500 cycles (Taken reference Efeoglu 2005)

Cross-References

- [Doped MoS₂ Coatings and Their Tribology](#)
- [MoS_x Coatings by Closed-Field Magnetron Sputtering](#)
- [Solid Lubricants](#)
- [Solid Lubricants for Space Mechanisms](#)
- [Thin Film Lubrication](#)

References

- E. Arslan, F. Bulbul, I. Efeoglu, The structural and tribological properties of MoS₂-Ti composite solid lubricants. *Tribol. Trans.* **47**, 218–226 (2004)
- E. Arslan, F. Bulbul, I. Efeoglu, The effect of deposition parameters and Ti content on structural and wear properties of MoS₂-Ti coating. *Wear* **259**, 814–819 (2005)
- E. Arslan, O. Baran, I. Efeoglu, Y. Totik, Evaluation of adhesion and fatigue of MoS₂-Nb solid-lubricant films deposited by pulsed-dc magnetron sputtering. *Surf. Coat. Technol.* **202**, 2344–2348 (2008)
- E. Arslan, Y. Totik, O. Bayrak, I. Efeoglu, A. Celik, High temperature friction and wear behavior of MoS₂/Nb coating in ambient air. *J. Coat. Technol. Rev.* **7**(1), 131–137 (2010)
- B. Bhushan, B.K. Gupta, *Handbook of Tribology* (McGraw-Hill, New York, 1991)
- F. Bulbul, I. Efeoglu, E. Arslan, The effect of bias voltage and working pressure on S/Mo ratio at MoS₂-Ti composite films. *Appl. Surf. Sci.* **253**, 4415–4419 (2007)
- S.V. Didziulis, P.D. Fleischauer, B.L. Soriano, M.N. Gardos, Chemical and tribological studies of MoS₂ films on SiC substrates. *Surf. Coat. Technol.* **43–44**, 652–662 (1990)
- W.M.R. Divigalpitiya, R.F. Frindt, S.R. Morrison, Effect of humidity on spread NbS₂films. *J. Phys. D: Appl. Phys.* **23**, 966 (1990)
- C. Donet, A. Erdemir, Solid lubricant coatings: recent developments and future trends. *Tribol. Lett.* **17**(3), 389–397 (2004)
- I. Efeoglu, Co-sputtered Mo:S:C:Ti:B based coating for tribological applications. *Surf. Coat. Technol.* **200**, 1724–1730 (2005)
- I. Efeoglu, Deposition and characterization of a multilayered-composite solid lubricant coating. *Rev. Adv. Mater. Sci.* **14**, 14–34 (2007)
- I. Efeoglu, F. Bulbul, Effect of crystallographic orientation on the friction and wear properties of Mo_xS_y-Ti coatings by pulsed-dc in nitrogen and humid air. *Wear* **258**, 852–860 (2005)
- I. Efeoglu, O. Baran, F. Yetim, S. Altintas Tribological, Characteristics of MoS₂-Nb solid lubricants films in different tribo-test conditions. *Surf. Coat. Technol.* **203**, 766–770 (2008)
- I. Efeoglu, S. Altintas, O. Baran, D. Ugur, Tribological characteristics of MoS₂-Nb solid lubricants films in different tribo-test conditions, TUBITAK Report in Turkish pp. 38–89, 2009a
- I. Efeoglu, J. Hardell, B. Prakash, *Interfacial scratch adhesion behaviour of multilayered TiB₂/TiBN-MoS₂/MoS₂-Ti Based PVD Coatings, ICMCTF09* (USA, San Diego, 2009b)
- P.D. Fleischauer, Fundamental Aspects of the Electronic Structure, Performance of Sputtered MoS₂ Films. *Thin Solid Films* **154**, 309–322 (1987)
- W.E. Jaminson, Electronic effects on the lubricating properties of molybdenum disulfide, in *Proceedings, 2nd International Conference on Solid Lubrication*, Denver, CO, p. 15, 1978
- A.R. Lansdown, *Molybdenum Disulphide Lubrication*. Tribology Series, vol. 35 (Elsevier, Amsterdam, 1999), p. 21
- D.-K. Lee, S.-H. Lee, J.-J. Lee, The structure and mechanical properties of multilayer TiN/(Ti_{0.5}Al_{0.5})N coatings deposited by plasma enhanced chemical vapor deposition. *Surf. Coat. Technol.* **169–170**, 433–437 (2003)
- K. Miyoshi, Solid lubricant fundamental and applications (NASA/TM-1998-107249/CHIREV1)
- N.M. Renevier, V.C. Fox, D.G. Teer, J. Hampshire, Coatings characteristics and tribological properties of sputter-deposited MoS₂/metal composite coatings deposited by closed field unbalanced magnetron sputter ion plating. *Surf. Coat. Technol.* **127**, 24–37 (2000)
- N.M. Renevier, J. Hampshire, V.C. Fox, J. Witts, T. Allen, D.G. Teer, Advantages of using self-lubricating, hard, wear-resistant MoS₂-based coatings. *Surf. Coat. Technol.* **142**, 67–77 (2001)
- T. Spalvis, *Lubrication with sputtered MoS₂ films: Principles, operation limitations*. NASA Technical Memorandum, vol. 105292 (Lewis Research Center Cleveland, Ohio, USA, 1991)
- B.C. Stupp, Synergistic effects of metals co-sputtered with MoS₂. *Thin Solid Films* **84**, 257–267 (1981)
- D.G. Teer, New solid lubricant coatings. *Wear* **251**, 1068–1074 (2001)
- G. Weise, A. Teresiak, I. Bacher, P. Markschlager, G. Kampschulte, Influence of magnetron sputtering process parameters on wear, properties of steel /Cr₃Si or Cr/MoS_x. *Surf. Coat. Technol.* **76–77**, 382–392 (1995)

Squats

- [Rolling Contact Fatigue \(RCF\)](#)

Squeeze Film Bearing Damper

- [Squeeze Film Dampers](#)

Squeeze Film Dampers

LUIS SAN ANDRÉS

Department of Mechanical Engineering, Turbomachinery Laboratory, Texas A&M University, College Station, TX, USA

Synonyms

[Squeeze film bearing damper](#)

Definition

Squeeze film bearing dampers are lubricated elements providing viscous damping in mechanical systems. Squeeze film dampers in rotating machinery provide structural isolation, reduce the amplitudes of rotor response to imbalance, and, in some instances, assist in suppressing rotordynamic instability.

Background

The most commonly recurring problems in rotordynamics are excessive steady state synchronous vibration levels and subsynchronous rotor instabilities. The first problem may be reduced by improved balancing, or by introducing modifications into the rotor-bearing system to move the system-critical speeds out of the operating range, or by introducing external damping to limit peak amplitudes at traversed critical speeds. Subsynchronous rotor instabilities may be avoided by eliminating the instability mechanism, by raising the natural frequency of the rotor-bearing system as high as possible, or by introducing damping to increase the onset rotor speed of instability (Vance 1988; Childs 1993).

Lightweight, high-performance engines exhibit a trend towards increased flexibility leading to a high sensitivity to imbalance with large vibration levels and reduced reliability. Squeeze film dampers (SFDs) are essential components of high-speed turbomachinery since they offer the unique advantages of dissipation of vibration energy and isolation of structural components, as well as the capability to improve the dynamic stability characteristics of inherently unstable rotor-bearing systems. SFDs are used primarily in aircraft jet engines to provide viscous damping to rolling element bearings, which themselves have little or no damping. One other important application is related to high-performance compressor units where SFDs are installed in series with tilting pad bearings to reduce (soften) bearing support stiffness while providing additional damping as a safety mechanism to prevent rotordynamic instabilities. In addition, in geared compressors, the SFD assists in reducing and isolating multiple frequency excitations transmitted through the bull gear, for example (San Andrés 2010).

Zeidan et al. (1996) give a history of the SFD in jet engines and detail design practices for successful SFD operation in commercial turbomachinery. Adilleta and Della Pietra and Adilleta (2002) provide a comprehensive review of the relevant analytical and experimental work conducted on SFDs. San Andrés and Delgado (2007) discuss more recent SFD experimental research and present a mechanically sealed SFD impervious to air entrainment.

In spite of the many successful applications, industry often recognizes that the design of SFDs is based on overly simplified predictive models that either fail to incorporate or simply neglect unique features (structural and fluidic) that affect the damper dynamic force performance. Actual damper performance can range from erratic to nonfunctioning, depending on the operating conditions.

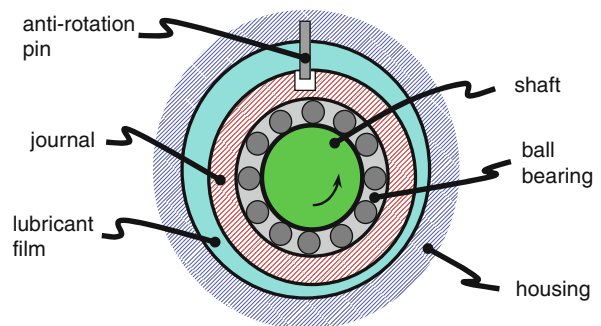
Issues such as lubricant cavitation or air entrainment are of fundamental interest (San Andrés and Diaz 2003).

Key Applications

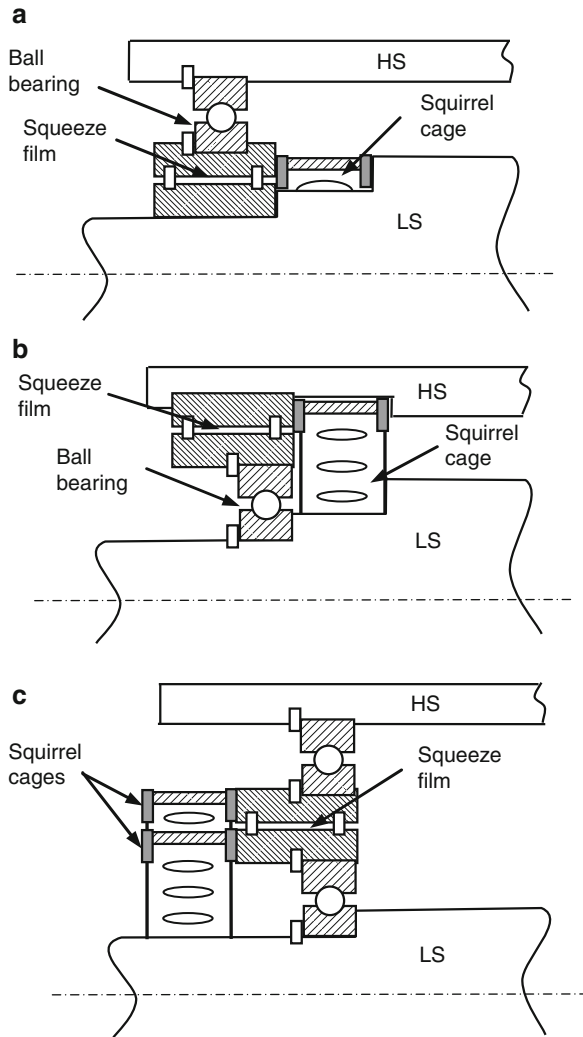
Figure 1 shows a typical SFD configuration consisting of an inner nonrotating journal and a stationary outer bearing, both nearly equal in diameter. The journal is mounted on the external race of a rolling element bearing and prevented from spinning with loose pins or a squirrel cage that provides a centering elastic mechanism. The annular squeeze film, typically less than 0.250 mm, between the journal and housing is filled with a lubricant provided as a splash from the rolling element bearing lubrication system or by a dedicated pressurized delivery. In operation, as the journal moves due to dynamic forces acting on the system, the fluid is displaced to accommodate these motions. As a result, hydrodynamic squeeze film pressures exert reaction forces on the journal and provide for a mechanism to attenuate transmitted forces and to reduce the rotor amplitude of motion.

Figure 2 shows conceptual views of intershaft dampers for multiple-spool gas turbine engines. These dampers are subject to whirl motions resulting from the combined imbalance response of both low-speed (LS) and high-speed (HS) rotors. Most SFDs in US aircraft engines incorporate the arrangements in Fig. 2a and b where the journal (and rolling element bearing) is elastically supported, and the bearing is rigidly attached to the engine frame. The (soft) spring support and squeeze film damper “see” the same deflections though the dynamic loads divide *unequally* between them.

Dampers in jet engines operate with low levels of pressurization (2 or 3 bar) to avoid excessive weight and volume in the lubrication system. Note also that most aircraft engines do not use any type of hydrodynamic journal bearings to avoid the risk of fluid film



Squeeze Film Dampers, Fig. 1 Typical squeeze film damper (SFD) configuration



Squeeze Film Dampers, Fig. 2 Schematic views of intershaft damper configurations: **(a)** squeeze film rotates with low-speed (LS) rotor, **(b)** squeeze film rotates with high-speed (HS) rotor, **(c)** double ball bearing-squirrel cage design

bearing-induced instabilities. (However, the inter-spool fluid film bearing in some dual shaft jet engines is known to be a source of such instabilities, see Fig. 2a.)

The amount of damping produced is the critical design consideration. If damping is too large, the SFD acts as a rigid constraint to the rotor-bearing system with large forces transmitted to the supporting structure. If damping is too light, the damper is ineffective and likely to permit large amplitudes of vibratory motion with likely subsynchronous motions. Note also that a damping element to be effective needs to be "soft," thus allowing for

motion at the location of the support, in particular for the modes of vibration of interest.

The damper geometry (length, diameter, and clearance), operating speed, and fluid properties (density and viscosity) determine, on first instance, the dynamic forced performance of SFDs. However, there are other important considerations that ultimately determine an appropriate operation. The relevant issues are:

- Kinematics of journal (tied to rotor system and acting forces)
- Level of supply pressure for adequate flow rate and cooling
- Feeding and end sealing mechanisms
- Fluid inertia effects
- Type of lubricant dynamic cavitation (vapor or gaseous) or air ingestion and entrapment

Scientific Fundamentals

Most dampers in practice are of short axial length, $L/D < 0.50$, and accommodate some type of end seals to increase their damping capability. SFDs include additional features such as high-resistance orifices for pressure delivery and discharge and/or deep grooves acting as flow sources or sinks of uniform pressure.

Squeeze film damper reaction forces and force coefficients are conveniently divided into two major types related to the specific journal center kinematics. For imbalance response analyses, SFD forces are obtained under the assumption of circular centered orbits. The model is applicable when the rotor traverses a critical speed, for example, where the imbalance force induces large amplitude orbital motions as the system may have little damping. On the other hand, for rotordynamic critical speed and stability analyses, SFD force coefficients are obtained for small amplitude journal center motions about a static (equilibrium) position. Only recently have computational tools analyzed rotor-bearing system transient response events by considering the instantaneous SFD reaction forces as a function of the time-varying journal kinematics that satisfy the equations of motion of the rotating system.

Figure 3 depicts a schematic view of a journal whirling within its bearing of radius R ($1/2$ diameter D) and length L . Lubricant of density ρ and viscosity μ fills the radial clearance c between the bearing and its journal. The film thickness h is squeezed as the journal whirls and displaces fluid. The film thickness equals

$$h = c + (e_{x_o} + \Delta e_x(t)) \cos(\Theta) + (e_{y_o} + \Delta e_y(t)) \sin(\Theta) \quad (1)$$

Squeeze Film Dampers, Table 1 Linearized force coefficients for open ends SFD (small amplitude motions about journal off-centered position)

Full film model (no cavitation)	π -film model (cavitated)
$C_{XX} = \mu D \left(\frac{L}{c}\right)^3 \frac{\pi(1+2\epsilon^2)}{2(1-\epsilon^2)^2}$	$C_{XX} = \mu D \left(\frac{L}{c}\right)^3 \frac{\pi}{2} \left[\frac{3\epsilon + \frac{1+2\epsilon^2}{2}}{2(1-\epsilon^2)^2} \right]$
$C_{XY} = 0$	$C_{XY} = \mu D \left(\frac{L}{c}\right)^3 \frac{\epsilon}{(1-\epsilon^2)^2}$
$C_{YY} = \mu D \left(\frac{L}{c}\right)^3 \frac{\pi}{2(1-\epsilon^2)^{3/2}}$	$C_{YY} = \mu D \left(\frac{L}{c}\right)^3 \frac{\pi}{4(1-\epsilon^2)^{3/2}}$
$C_{YX} = 0$	$C_{YX} = 0$
$M_{XX} = \rho D \left(\frac{L^3}{c}\right) \frac{\alpha\pi[1-(1-\epsilon^2)^{1/2}]}{12\epsilon^2(1-\epsilon^2)^{1/2}}$	$M_{XX} = \rho D \left(\frac{L^3}{c}\right) \frac{\alpha(i-\pi-2\epsilon)}{24\epsilon^2}$
$M_{XY} = 0$	$M_{XY} = \rho D \left(\frac{L^3}{c}\right) \frac{\alpha \left[\ln \left\{ \frac{(1-\epsilon)}{(1+\epsilon)} \right\} - 2\epsilon \right]}{24\epsilon^2}$
$M_{YY} = \rho D \left(\frac{L^3}{c}\right) \frac{\alpha\pi[1-(1-\epsilon^2)^{1/2}]}{12\epsilon^2}$	$M_{YY} = \rho D \left(\frac{L^3}{c}\right) \frac{\alpha\pi[1-(1-\epsilon^2)^{1/2}]}{24\epsilon^2}$
$M_{YX} = 0$	$M_{YX} = 0$

$$\begin{bmatrix} F_X \\ F_Y \end{bmatrix} = - \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} \begin{bmatrix} V_X \\ V_Y \end{bmatrix} - \begin{bmatrix} M_{XX} & M_{XY} \\ M_{YX} & M_{YY} \end{bmatrix} \begin{bmatrix} A_X \\ A_Y \end{bmatrix} \quad (3)$$

where (V_X, V_Y) and (A_X, A_Y) are the *instantaneous* journal center velocities and accelerations in the X and Y directions, respectively. $(C_{\alpha\beta}, M_{\alpha\beta})_{\alpha,\beta=X,Y}$ are the damping and inertia force coefficients, respectively. Recall that a SFD does not produce direct stiffnesses, that is, without journal spinning, a damper cannot generate film pressures given a journal static displacement.

Table 1 shows formulas for the linearized force coefficients of a short length, open ends SFD. The coefficients are nonlinear functions of the static journal eccentricity ratio $(\epsilon = e_s/c)$. The fluid inertia or added mass coefficients are strictly valid for small-to-moderate squeeze film Reynolds numbers, $Re_s = \frac{(\rho \omega c^2)}{\mu} < 10$.

$i = \frac{2\cos(\epsilon)}{(1-\epsilon^2)^{1/2}}$, and $\alpha = 1.2$ – 1.0 for small to moderately large squeeze film Reynolds numbers ($Re_s < 50$). Note that the coefficients C_{YX} and M_{YX} are nil.

SFD Force Coefficients for Circular Centered Orbits

Figure 5 shows a SFD journal describing circular centered orbits of amplitude (e) and whirl frequency (ω) . The damper generates a constant reaction film force in a reference frame rotating with frequency ω . The radial

(F_r) and tangential (F_t) components of the damper reaction force are

$$F_r = -\{C_{rt} V_t + M_{rr} A_r\}; \quad F_t = -\{C_{tt} V_t + M_{tr} A_r\} \quad (4)$$

where $V_t = e\omega$ and $A_r = -e\omega^2$ are the journal center tangential speed and radial acceleration, respectively. (C_{tt} , C_{rt}) denote the direct and cross-coupled viscous damping coefficients, and (M_{rr} , M_{tr}) are fluid inertia force coefficients, respectively. Recall that SFDs *do not* generate stiffness coefficients (i.e., reaction forces due to static journal displacements). The archival literature misleads the designer when referring to a damper direct radial stiffness, $K_{rr} = C_{rr} \omega$, that is frequency dependent.

For the short-length open ends SFD model, the force coefficients using the rather simplistic π -film assumption (i.e., half the damper circumference develops film cavitation) are (Vance 1988)

$$C_{tt} = \frac{\pi \mu D}{4(1-\epsilon^2)^{3/2}} \left(\frac{L}{c}\right)^3; \quad C_{rt} = \frac{\mu \epsilon D}{(1-\epsilon^2)^2} \left(\frac{L}{c}\right)^3 \quad (5)$$

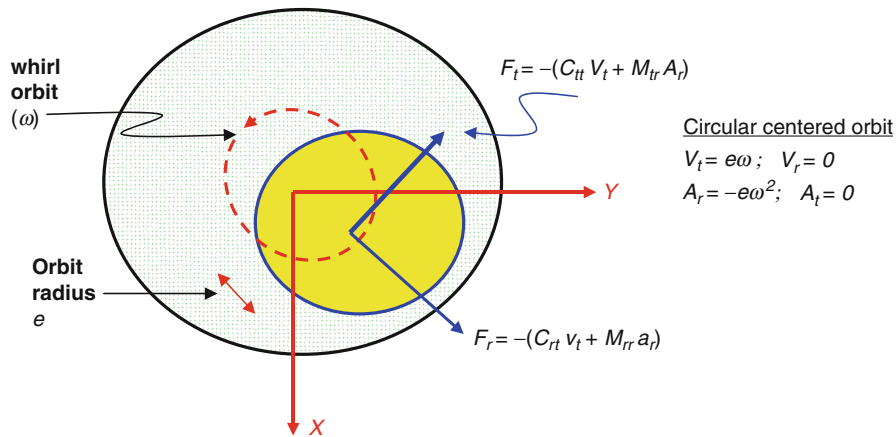
$$M_{rr} = \frac{\pi \rho D}{24} \left(\frac{L^3}{c}\right) \left[1 - 2(1-\epsilon^2)^{1/2}\right] \left\{ \frac{(1-\epsilon^2)^{1/2} - 1}{\epsilon^2(1-\epsilon^2)^{1/2}} \right\};$$

$$M_{tr} = -\frac{27}{140\epsilon} \rho D \left(\frac{L^3}{c}\right) \left[2 + \frac{1}{\epsilon} \ln\left(\frac{1-\epsilon}{1+\epsilon}\right)\right] \quad (6)$$

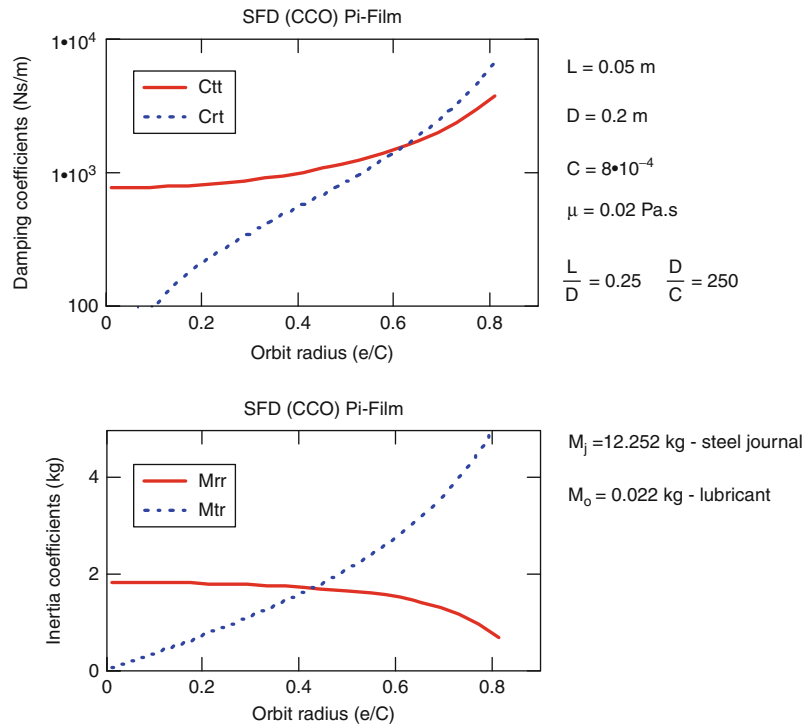
where (L, D, c) denote the damper axial length, diameter, and radial clearance, respectively, (μ, ρ) are the (effective) lubricant viscosity and density, and $\epsilon = e/c$ is the dimensionless orbit radius. The orbit radius (e) should not be confused with the static journal offset displacement, null in this case. Note that the coefficients in (5) and (6) are not strictly rotordynamic coefficients as their classical definition implies small amplitude motions (perturbations) about a journal equilibrium position.

The coefficients above are determined under the assumptions of an isoviscous and incompressible lubricant that is supplied with a low feed external pressure. Most importantly, the model assumes a squeeze film *fully submerged in a lubricant bath*. For the full film model (no oil cavitation), the direct coefficients (C_{tt} , M_{rr}) are twice the values given by (5) and (6), while the cross-coupled coefficients (C_{rt} , M_{tr}) are null. The inertia coefficients are strictly valid for small to moderate squeeze film Reynolds numbers, $Re_s = \frac{(\rho \omega c^2)}{\mu} < 10$.

Figure 6 depicts the damping and inertia force coefficients for a short length, open ends SFD describing circular centered orbits (CCOs). The damper length $L = 50$ mm,



Squeeze Film Dampers, Fig. 5 SFD model: circular centered orbit with radius e



Squeeze Film Dampers, Fig. 6 Open ends SFD force coefficients for circular centered motions (Short length π film model)

$c = 0.080$ mm, $L/D = 0.25$, with lubricant viscosity and density (μ, ρ) equal to 0.020 Pa.s and 890 kg/m³, respectively. The predicted force coefficients are highly nonlinear functions of the orbit radius (e). Note the large magnitudes of direct damping (C_{tt}) even for the centered position ($e = 0$). The rapid growth of the cross-coupled damping coefficient (C_{rt}) is referred as a “stiffness

hardening effect,” and the culprit of severe nonlinear (multiple valued) rotor response accompanied with jump-phenomenon and orbit-instability. However, these effects, mostly predicted by overly simplified theoretical analyses, are hardly ever reported in practice. Note that air entrainment is most prevalent for large amplitude orbital motions ($e \rightarrow c$) and high frequencies of operation,

determining a damper forced response quite different from the one derived from the force coefficients shown in Fig. 6.

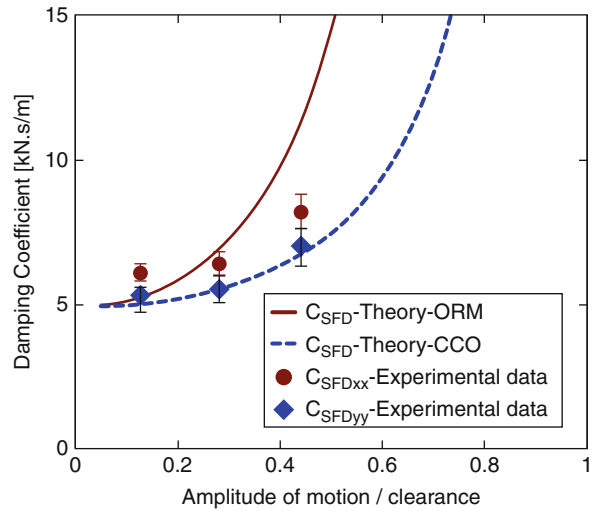
In actuality, stiffness hardening, $K_{rr} > 0$, is most likely due to contact and rubbing of the journal and bearing surfaces that may occur while a rotor traverses a critical speed with large orbital motions due to little damping or excessive rotor imbalance, for example. In these events, damper forces are negligible since the fluid film is probably ruptured with large amounts of air entrainment. Thus, predictions from a nonlinear rotor-SFD model based on the π -film short length SFD bear little relationship to reality.

The inertia force coefficients (M_{rr} , M_{rt}) have an effect on the system rotordynamic response. This is so in spite that the fluid mass, M_o , contained in the film ($\rho\pi D L c$) is just a few grams. Note that the journal mass ($M_j = \rho_s \pi R^2 L$) for a steel construction ($\rho_s = 7,800 \text{ kg/m}^3$) is 12.25 kg. Thus, the SFD added mass coefficients are of the same order of magnitude as the actual journal mass. Hence, fluid inertia in SFDs impacts the location of critical speeds in compact rotors operating at high rotational speeds.

There is good correlation between test derived and predicted force coefficients for SFDs operating with circular centered orbits (San Andrés 1996). The test damping coefficients (C_{td} , C_{rt}) fall in between the π - and full-film predictions. At low frequencies, the cavitation zone does not extend over half the damper circumference, and thus the damping coefficients approach the full film predictions. On the other hand, as the whirl frequency increases so does the squeeze film pressure and the cavitation zone extends. The experimental values thus approach those derived for the π -film model. It is most important to note that the experiments were conducted in a damper fully submerged within a lubricant bath. The test rig had closed any path that would permit the natural ingestion and entrapment of air. This condition in practice is most difficult to achieve. San Andrés and Delgado (2007) present more recent SFD parameters agreeing well with predictions, in particular for added mass coefficients (see Fig. 7).

The effect of air entrainment on squeeze film pressures and damper reaction forces has been thoroughly researched qualitatively and quantitatively in the last decade. San Andrés and Diaz (2003) report fundamental experimental results and advance an analytical model for thus most prevalent operating condition (see below for a further discussion on this issue).

A brief discussion follows on other physical and operating conditions that affect the performance of squeeze film dampers.



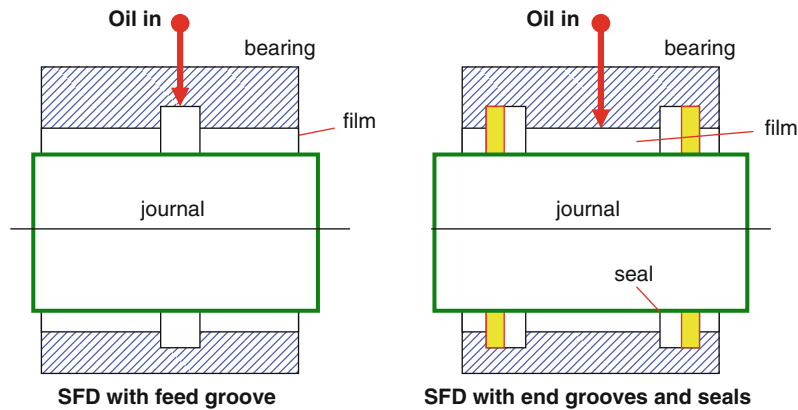
Squeeze Film Dampers, Fig. 7 Squeeze film damping coefficients identified from varying load amplitude – multi-frequency sine sweep forced excitations. Predictions for circular centered orbits (CCO) and radial motions about an off-centered journal static position (ORM)

SFDs with Feed Grooves

Some dampers are designed with feed and discharge grooves to ensure a continuous flow of lubricant through the squeeze film lands (see Fig. 8). A groove is thought to provide a uniform flow source with constant pressure around the bearing circumference. A central feed groove also divides the flow region into two separate squeeze film dampers working in parallel (i.e., the reaction forces from each land add).

For the central groove configuration, theory predicts forces about one-fourth less than that available for a damper with twice the land length and no groove. Experiments, however, demonstrate that grooved dampers generate much larger levels of forces than those derived from accepted theory. Large amplitude dynamic pressures are measured at the groove regions connecting the two squeeze film regions. Thus, a central groove does not isolate the adjacent film lands, but rather interacts with the squeeze film regions (Arauz and San Andres 1997; Childs et al. 2007).

Delgado (2008) presents a novel model for prediction of the forced response of grooved SFDs and grooved oil seal rings. The model includes fluid inertia and flow interactions at the groove-film land interface that amplify the generation of squeeze film pressures. Delgado's model predictions are in excellent agreement with measured stiffness, damping and inertia force coefficients in oil seal



Squeeze Film Dampers, Fig. 8 SFD: grooved configurations

rings with multiple cavities, and in dampers with inlet and discharge deep grooves.

SFDs with End Seals

SFDs usually incorporate some type of end seals to reduce the through flow and to amplify the viscous damping. The most common end seal configurations include O-rings, piston rings, and end plate (clearance gap) seals, as shown in Fig. 9. Measurements and analysis show larger forces for the end-sealed condition, although the lubricant heats rapidly (lower viscosity) for designs with little through flows. The design of end seals is highly empirical and requires leakage correction factors that can only be extracted from exhaustive experimentation. To date, only experience dictates the best type of sealing to be implemented.

SFDs in jet engine rotors incorporate piston rings as end seals. However, ring cocking and locking with a resulting excessive oil leakage is a pervasive problem. Implementing patented (proprietary) designs seems to resolve the reliability issue.

Many industrial compressor applications also implement dampers with O-ring end seals due to their simplicity and good sealing. However, these applications are restricted to low static loads and low temperatures. Material compatibility of the O-rings with the lubricant and gas external medium is a design consideration. Long-term relaxation and creep of the elastomeric O-rings, when supporting large static loads, is an issue usually overlooked that later can prove fatal.

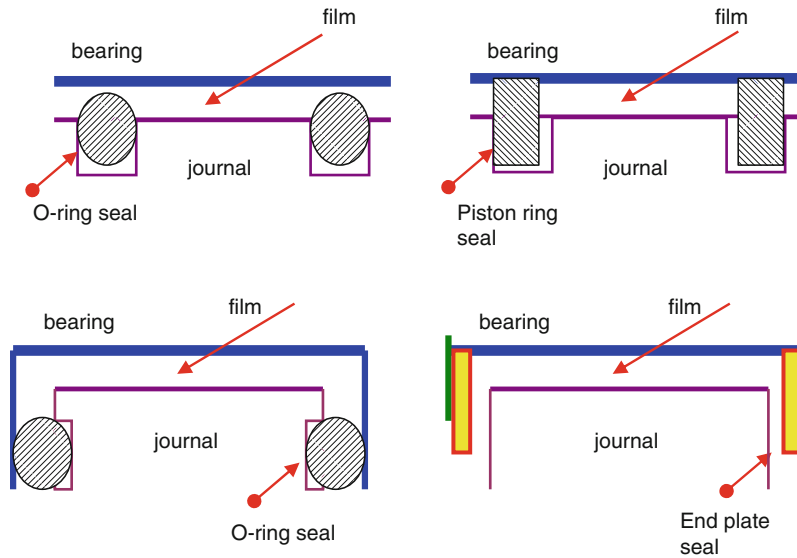
San Andrés and Delgado (2007) detail parameter identification measurements conducted on a squeeze film damper (SFD) featuring a non-rotating mechanical seal that effectively eliminates lubricant side leakage. The SFD-seal arrangement generates dissipative forces due to

viscous and dry-friction effects from the lubricant film and surfaces in contact. The identified system damping coefficients are frequency and motion amplitude dependent due to the dry friction interaction at the mechanical seal interface. Squeeze film force coefficients, damping and added mass, are in agreement with simple predictive formulas for an uncavitated lubricant condition and are similar for both flow restrictor sizes. The SFD-mechanical seal arrangement effectively prevents air ingestion and entrapment and generates predictable force coefficients for the range of frequencies tested.

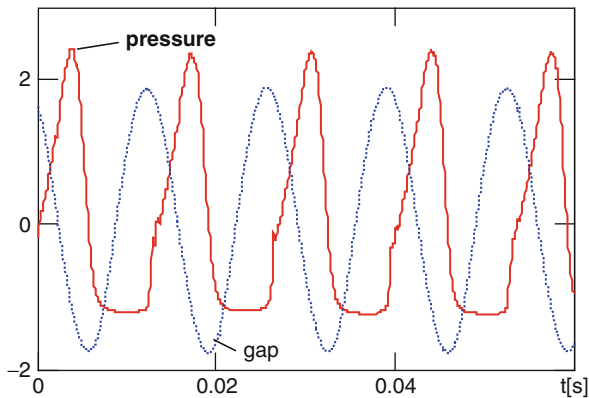
Lubricant Cavitation Versus Air Entrainment in SFDS

Zeidan et al. (1996) identify SFD operation with distinct types of dynamic fluid cavitation (vapor or gas) and a regime due to air ingestion and entrapment. The appearance of a particular condition depends on the damper type (sealed or open to ambient), magnitude of supply pressure and flow rate, whirl frequency, and magnitude of dynamic load producing (small or large) journal excursions within the film clearance.

Gas cavitation following the journal motion appears in ventilated (open ends) SFDs operating at low frequencies and with small to moderate journal amplitude motions. A well-defined cavitation bubble containing the release of dissolved gas in the lubricant or air entrained from the vented sides follows the whirling motion of the journal (i.e., the cavitation zone appears steady in a rotating frame). The traveling gas bubble appears not to affect the generation of the squeeze film pressure in the full film zone. The persistence of this cavitation regime upon reaching steady operating conditions (high frequencies) in an aircraft application is remote.



Squeeze Film Dampers, Fig. 9 SFD: types on end seals



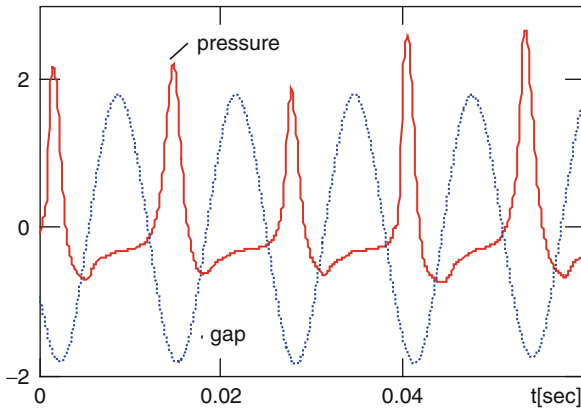
Squeeze Film Dampers, Fig. 10 Dynamic film pressure (bar) and local film gap (mm \times 10) in a flooded SFD leading to vapor cavitation

Lubricant vapor cavitation appears in dampers with tight end seals that prevent entrainment of the external gas media and for operation with a sufficiently large supply pressure. In this last case, the through oil flow also prevents the ingestion of air. Furthermore, the lubricant must be relatively free of dissolved gases such as air, a condition not readily found in practice.

Figure 10 depicts a measured dynamic film pressure versus time in a damper operating with lubricant vapor cavitation. The experiment illustrates the variation of dynamic squeeze film pressure and gap (film thickness) for five periods of journal orbital motion. The whirl

frequency and centered journal orbital amplitude equal 75 Hz and 0.180 mm, respectively. The damper radial clearance is 0.343 mm. The damper is fully flooded in a lubricant bath; the supply pressure is 1.45 bar and the discharge is at atmospheric pressure. Note that the pressure profile is smooth and shows nearly identical shapes for each consecutive period of motion. A (flat) constant pressure zone develops at nearly zero absolute pressure, and it corresponds to the rupture of the film and formation of a vapor-filled cavity. The cavity appears only during that portion of the journal motion cycle where the film gap increases. The vapor bubble collapses immediately as the local pressure rises above the lubricant vapor pressure. In general, correlations of measured pressures and vapor cavitation extent with predictions based on traditional film rupture models are satisfactory (Diaz and San Andrés 1999).

Air ingestion and entrapment appear in vented dampers operating at high frequencies and with low magnitudes of supply (feed) pressure (i.e., small throughout flow rates). Figure 11 depicts the measured dynamic film pressure versus time in a SFD with air entrapment. The operating conditions are identical to those for the measurements depicted in Fig. 10, except that the damper is open to ambient conditions (i.e., not submerged in an oil bath). A suction pressure draws air into the thin film at the locations where the local film gap is increasing. The cyclic fluid motion leads to air entrapment, with bubbles remaining in the zones of dynamic pressure generation above ambient. Air ingestion leads to the formation of



Squeeze Film Dampers, Fig. 11 Dynamic film pressures (bar) and local film gap (mm \times 10) in a SFD operating with air entrainment

intermittent air fingering surrounded by liquid striations (see Fig. 12 for details). These islands of air may shrink, break up into smaller zones, or diffuse within the lubricant. The size and concentration of the ingested air fingers depend on the journal whirl frequency and amplitude and the flow rate. The fluid at the damper discharge is cloudy and foamy (San Andrés and Diaz 2003).

The dynamic pressures with air entrainment, Fig. 11, show important differences when compared with those pressures induced by lubricant vapor cavitation (Fig. 10). In the case of air ingestion, the squeeze film pressures differ markedly from one period to the next, and with peak pressures showing large variations. Furthermore, the pressure flat zone is nearly at ambient pressure. Note that subambient film pressures are also generated.

The vast majority of SFDs inevitably operate with foam-like fluids considering the low values of pressure supply (small flow rate), large damper clearances, and high operating whirl frequencies. Of course, mixed operation regimes can also occur in practice. For instance, tightly sealed dampers may show both vapor and air entrainment type cavitation where gas bubbles may coexist around a large lubricant vapor bubble. Note that the entrapment of air delays the increase of film pressures since there is less liquid lubricant filling the damper clearance. Ultimately, operation at high frequencies leads to an increase in air ingestion, preventing any further oil vapor cavitation, and reducing considerably the forces available from the SFD.

Careful experimentation demonstrates that air ingestion and entrapment degrades considerably the forced response of open ends SFDs (San Andrés and Diaz 2003).

A simple criterion gives the likelihood of air entrainment in a damper. A feed-squeeze flow parameter (γ) relates the lubricant supply flow rate Q_{oil} to the dynamic change in volume within the squeeze film gap, that is,

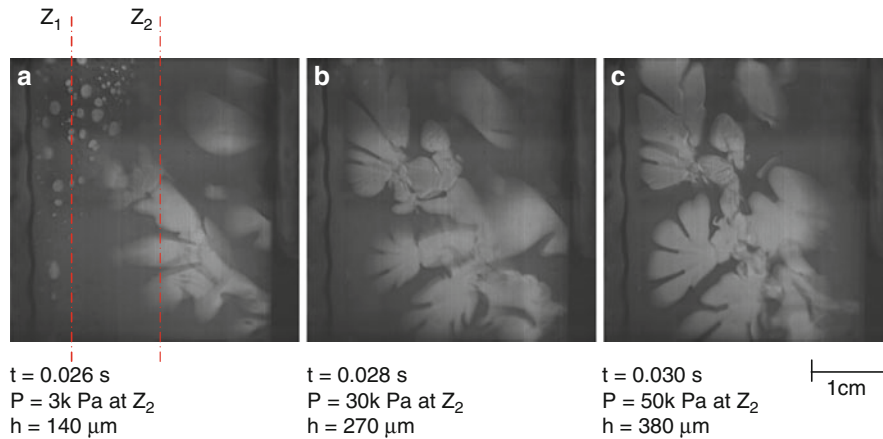
$$\gamma = \frac{Q_{oil}}{\pi DL \dot{e} \omega} \quad (7)$$

If $\gamma > 1$ then no air entrainment occurs, that is, the through flow is sufficient to fill the volume change caused by the journal whirl motion. On the other hand, air ingestion and entrapment will occur when $\gamma < 1$. The lower the feed-squeeze parameter (γ), the more severe the degradation in damper forced performance. The experimental results advance an empirical correlation between γ and the amount of air entrained (volume concentration of air) in the lubricant, thus providing certainty in the modeling of the mixture. Note that Q_{oil} is proportional to the difference between lubricant supply pressure and discharge pressure and to the flow conductances in the film lands and through the feed ports. The flow conductances ($\sim 1/\text{resistances}$) are a function of the damper clearance and feed characteristics, lubricant and mixture viscosities, and so on. Thus, air entrainment is device dependent, and its severity increases with the amplitude and frequency of journal motion. Air ingestion can be prevented by increasing the supply pressure (not practical) to ensure a sufficiently large through lubricant flow rate.

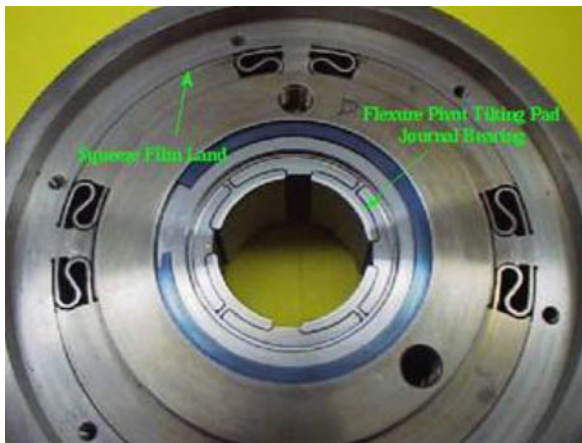
Modern Squeeze Film Dampers

The basic design of SFDs changed little until the late 1990s when the novel wire-EDM processes allowed the construction of integral SFDs, which offer distinct advantages such as reduced overall weight and length of the damper structure with less number of parts, accuracy of positioning (centering), and a split segment construction allowing easier assembly, inspection, and retrofit than with any other type of damper.

Flexure pivot tilting pad bearings offer similar construction features while minimizing (assembly) stack up tolerances and avoiding pivot wear and fretting. These features are most important in aircraft engines where reduced weight and size are of utmost consideration. The integral damper, as shown in Fig. 13, comprises segmented pads instead of a fully cylindrical journal. Thin, structured webs attach the inner and outer rings and perform the function of elastic supports. The thin gap between the pads and the outer ring forms the squeeze film lands. Each pad can be manufactured with a different clearance to counter the static deflection due to rotor weight. End seals restricting the axial flow through the



Squeeze Film Dampers, Fig. 12 Photographs of SFD flow field with air ingestion and entrapment. Tests with whirl frequency at 25 Hz and feed pressure 1.93 bar. Elapsed time for photographs is 2 ms (period = 40 ms)



Squeeze Film Dampers, Fig. 13 Series integral SFD and flexure pivot tilting pad bearing

film lands dampers provide the means to increase the damping coefficients by raising the hydrodynamic pressure in a pad film land. The series combination of a tilting pad bearing and a squeeze film damper has been implemented in numerous compressors to introduce flexibility and damping to the bearing support. The proper design of these two mechanical elements allows for the optimum damping coefficient at the bearing support and accurate relocation of the (rigid mode) rotor bearing system critical speeds away from the operating speed range. De Santiago et al. (1999) provide experimental verification and theoretical validations of the damping capability of sealed integral dampers and demonstrate

the benefits of this novel technology for application in modern high-performance turbomachinery.

Closure

Decades of practice demonstrate that SFDs generate the required damping even when operating with persistent air entrainment. Damper support flexibility (structural stiffness) is the key parameter that allows the device intended operation in a practical application. Incidentally, the actual reduction of predicted damping at high frequencies (due to air ingestion, for example) is beneficial in rotor-bearing systems operating at supercritical speeds. However, the trend toward higher operating speeds and more stringent operating conditions demands a reliable predictive physical model, experimentally verified.

Childs (1993) noted that, because of lubricant cavitation and unquantifiable air ingestion, correlation between theory and experiment is less compelling for SFDs than journal bearings. In practice SFDs operate with low magnitudes of oil feed pressure (5 bar max.) that generally do not prevent the lubricant in the fluid film lands from liquid vaporization or entrainment of external gaseous media into the film lands. Open-ends SFDs are prone to develop a flow regime where the ingestion of air leads to the formation of a *bubbly* lubricant. Actual practice demonstrates that air ingestion greatly affects the SFD dynamic forced response.

Digital movies obtained in a squeeze film damper operating with air entrainment are available online at <http://rotorlab.tamu.edu> (Research sponsored by US National Science Foundation, 1996–2000). The movies vividly depict air ingestion and entrapment cycles with

notable effects on the recorded squeeze film pressures and ensuing damper dynamic forced response.

Most (numerical) models for prediction of finite length SFD forced response assume lubricant vapor cavitation (i.e., an operating condition likely to be found with very tight end seals or if the damper is fully submerged in a lubricant bath). Understanding of air entrainment, a pervasive phenomenon in SFDs, has just begun.

Cross-References

- [Cavitation Phenomena and Numerical Analysis](#)
- [Lubricant Inertia and Its Effect on Lubrication](#)
- [Thin Film Lubrication](#)

References

- G. Arauz, L. San Andrés, Experimental force response of a grooved squeeze film damper. *Tribol. Int.* **30**, 77–86 (1997)
- D. Childs, *Turbomachinery Rotordynamics* (Wiley, New York, 1993)
- D.W. Childs, M. Graviss, L.E. Rodriguez, The influence of groove size on the static and rotordynamic characteristics of short, laminar-flow annular seals. *ASME J. Tribol.* **129**(2), 398–406 (2007)
- O. De Santiago, L. San Andrés, J. Oliveras, Imbalance response of a rotor supported on open-ends, integral squeeze film dampers. *ASME J. Gas Turbines Power* **121**(4), 718–724 (1999)
- A. Delgado, *A Linear Fluid Inertia Model for Improved Prediction of Force Coefficients in Grooved Squeeze Film Dampers and Grooved Oil Seal Rings*, Ph.D. Dissertation, Texas A&M University, College Station, 2008
- L. Della Pietra, G. Adiletta, The squeeze film damper over four decades of investigations. Part I: characteristics and operating features. *Shock Vib. Dig.* **34**(1), 3–26 (2002). Part II: rotordynamic analyses with rigid and flexible rotors. *Shock Vib. Dig.* **34**(2), 97–126 (2002)
- L. San Andrés, Theoretical and experimental comparisons for damping coefficients of a short length open-end squeeze film damper. *ASME J. Eng. Gas Turbines Power* **118**, 810–815 (1996)
- L. San Andrés, A. Delgado, Identification of force coefficients in a squeeze film damper with a mechanical end seal- centered circular orbit tests. *ASME J. Tribol.* **129**(3), 660–668 (2007)
- L. San Andrés, S. Diaz, Flow visualization and forces from a squeeze film damper with natural air entrainment. *ASME J. Tribol.* **125**, 325–333 (2003)
- L. San Andrés, *Modern Lubrication Theory, Squeeze Film Dampers*. Lecture notes, vol. 13 (2010), Texas A & M University Digital Libraries, <http://repository.tamu.edu/handle/1969.1/93197>. Accessed November 2010
- J. Vance, *Rotordynamics of Turbomachinery* (Wiley, New York, 1988)
- E.L. Zeidan, L. San Andrés, J. Vance, Design and application of squeeze film dampers in rotating machinery, in *Proceedings of the 25th Turbomachinery Symposium*, Houston, 1996, pp. 169–188

Squeeze Film Gas Bearing

- [Squeeze Film Gas Lubrication](#)

Squeeze Film Gas Lubrication

CODA H. PAN

Global Technology, Millbury, MA, USA

Synonyms

[Compressible squeeze film](#); [Gaseous squeeze film](#); [Squeeze film gas bearing](#); [Squeezed gas film](#)

Definition

Generation of pressure in a thin gas film by high-frequency normal oscillation to maintain separation of its enclosing surfaces.

Scientific Fundamentals

As a means of reducing friction between parallel surfaces, squeeze film gas bearings deserve the distinction of being one of the youngest entries into the world of tribology.

Historical Events

First knowledge of the phenomenon was disclosed in 1957 by Sir G. I. Taylor, to explain an unexpected discovery of elevated pressure at the center of a high-speed disk in close proximity with a fixed “parallel” wall. Sir Taylor reasoned the finding to be a low Reynolds number compressible flow phenomenon due to an inadvertent non-perpendicular attachment of the disk to the rotating shaft. He supported his explanation with a simple analysis that can be regarded as the forerunner of the asymptotic theory of gaseous squeeze films.

The use of the term “squeeze film” began in an article by William A. Gross in *International Science and Technology* in 1963, identifying squeeze film as one of three mechanisms to achieve elevated pressure in a thin gas film. In the following year, a full-length technical article authored by E. O. J. Salbu was published.

Mechanism

Salbu performed his squeeze film bearing experiment using a circular disk mounted on a loud speaker voice coil. He used numerical computation to find the weight load that can be supported by the circular squeeze film bearing. Salbu’s disclosure led to several years of sponsorship by NASA to make use of squeeze film bearing in the form of a low-friction gimbal in the navigation platform of a space vehicle. From the NASA program, asymptotic squeeze film theory was developed. The asymptotic squeeze film theory is very simple and easy to implement;

it is a practical alternative to direct numerical computation for engineering studies in squeeze film applications.

The squeeze film bearing depends on a transverse oscillation, at amplitude that is a moderate fraction of the nominal clearance (► [Mathematical Foundation of Fluid Lubrication Theory](#)), which is sustained at a high squeeze number defined as

$$\sigma \equiv 12\mu(2\pi f_D)(p_a C^2)^{-1} R^2$$

At startup, when the film gap is contracting quickly, viscous friction retards outward flow at the edges; as the squeeze motion reverses direction, flow entry from ambient fills the enlarging gap with little resistance. The inflow surplus results in a net gain of gas mass to increase the mean film pressure so that inflow surplus during the next squeeze cycle would be reduced by the higher mean pressure. Inflow surplus vanishes when the mean film pressure reaches a level that increases with the excursion amplitude and the squeeze frequency; the level of squeeze pressurization levels off as $\sigma \rightarrow \infty$. Typically squeeze film pressurization becomes significant at a high frequency, although it is a low Reynolds number gas flow phenomenon that is not directly related to acoustics. Nevertheless, it is desirable to make the operating frequency ultrasonic to avoid the presence of undesirable audible sound.

A comparison of Salbu's numerical computation and the asymptotic theory is shown below ([Fig. 1](#)).

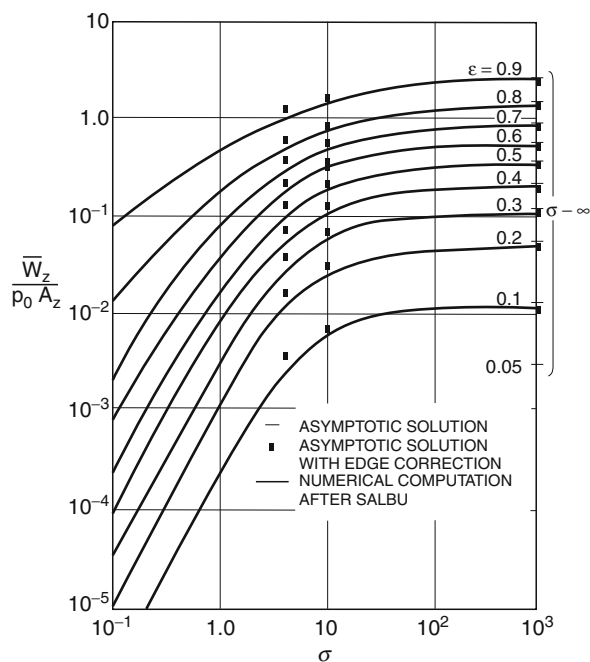
Ambient Humidity

A squeeze film is pressurized by drawing in air from the ambient while maintaining thermal equilibrium. If the ambient humidity is high, psychrometric equilibrium of the humid air in the pressurized interior of the squeeze film can result in condensation of aqueous liquid phase to cause squeeze film failure. Therefore, humidity control of the operating environment should not be overlooked in the deployment of a squeeze film device.

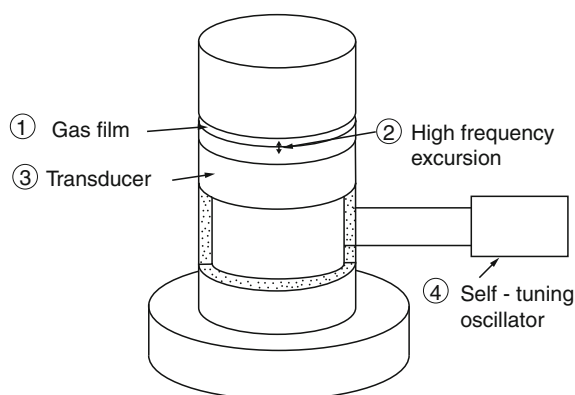
The Total System

The principal components of a squeeze film bearing are the bearing structure, an actuator, and its power supply. [Figure 2](#) illustrates the possibilities for assembling a system in forms ranging from a laboratory demonstration to a “field-ready” package.

Piezoelectric transduction is the preferred method for maintaining the desired squeeze motion in view of the general availability of the suitable material and the abundance of industrial experience of ultrasonic applications. For efficient operation, actuation is tuned to a resonant frequency of the bearing-transducer structure



Squeeze Film Gas Lubrication, Fig. 1 Circular squeeze film characteristics (Pan 1967, 1970)

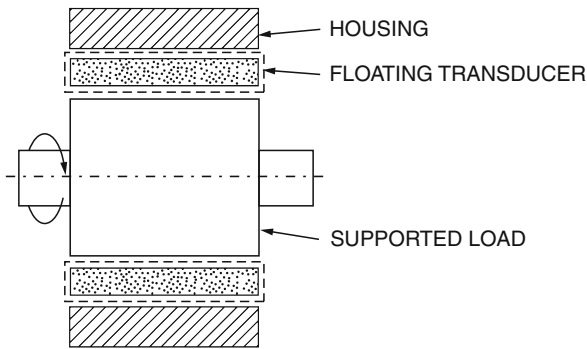


Squeeze Film Gas Lubrication, Fig. 2 Elements of a squeeze film bearing

and a self-tuning drive circuit is used to adjust the drive frequency to compensate for temperature-dependent shift of the resonant frequency.

Transducer Designs

[Figure 2](#) illustrates a possible laboratory set-up, using a piezoelectric tube in lieu of a loud speaker voice coil for actuation of oscillating squeeze excursions of the circular disk, to reproduce Salbu's experiment.



Squeeze Film Gas Lubrication, Fig. 3 Self-centering floating-sleeve squeeze film journal bearing (Pan et al. 1966)

Figure 3 illustrates the deployment of a piezoelectric tube to function as a floating-sleeve journal bearing that is squeeze-actuated in a hoop resonance; due to the Poisson ratio effect, there is enough induced axial oscillation to achieve non-contact centering between the caps of the housing.

In these two examples, the excursion amplitude of the squeeze film is directly coupled to the longitudinal strain of the piezoelectric material, which saturates at about 10^{-4} . This limitation can be overcome by using the piezoelectric material as the driver of a composite transducer structure that can sustain a large squeeze motion without causing strain saturation of the piezoelectric material.

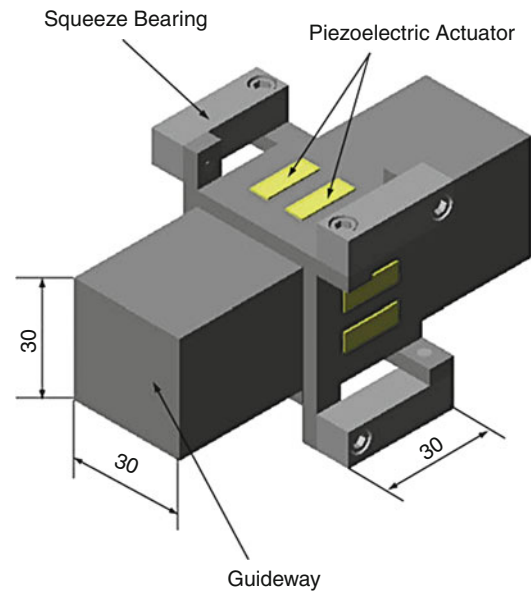
A composite transducer for a journal bearing-type squeeze film can be constructed with a thin-wall cylindrical shell with a piezoelectric ring attached near a nodal circle of a complex hoop mode resonance of the oscillating shell. Such a transducer can be used as a squeeze-film wire guide or an actuator slide.

A precision guideway, shown in Fig. 4, features composite squeeze film panels that are actuated by adhesive-fastened piezoelectric strips in a bending mode; it has been evaluated and shows much promise.

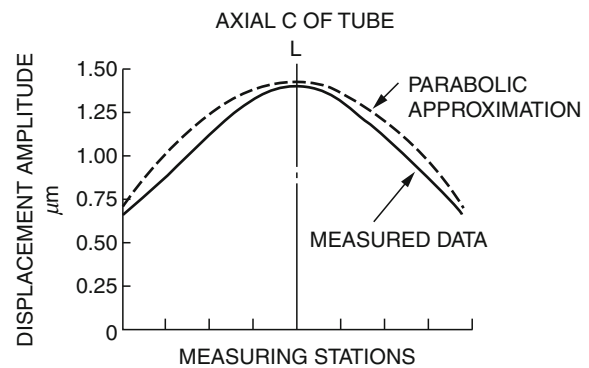
The excursion amplitude of the squeeze film bearing surface is not uniform in operation. The spatial profile of the excursion amplitude is needed for determination of the squeeze film pressure. The measured excursion profile of the floating sleeve journal bearing is seen to be approximately a parabola, as shown in Fig. 5.

A similar measurement of the excursion profile of the guideway transducer panel was made and curve-fitted into a topographical plot, as shown in Fig. 6.

During the process of designing a new squeeze film bearing system, the excursion profile of a transducer can be estimated from the resonance mode shape of the transducer structure; its operation as a squeeze film bearing



Squeeze Film Gas Lubrication, Fig. 4 Linear guideway constructed of composite squeeze film panels (Kobayashi et al. 2007, Ono et al. 2009)



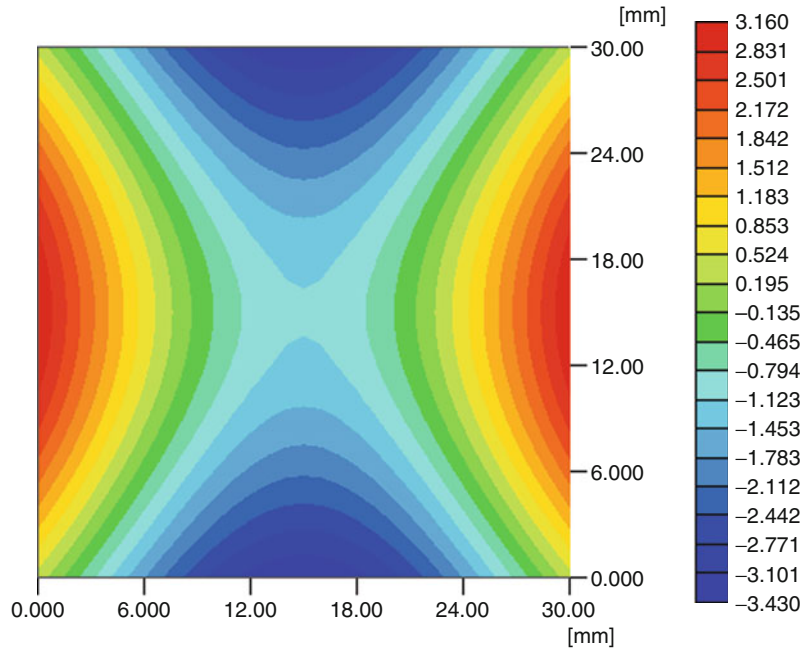
Squeeze Film Gas Lubrication, Fig. 5 Radial excursion amplitude of tubular transducer in hoop mode resonance (Pan et al. 1966)

would be assigned various amplitude levels in the design study as a parameter in performance prediction and selected for the specification of the drive electronics.

Squeeze Film Analysis

Components of Film Thickness

The squeeze film bearing distinguishes itself in the presence of a squeeze excursion profile $a_{\text{excursion}} \Delta \tilde{h} e^{i(2\pi f_0)t}$; $a_{\text{excursion}}$ is the excursion amplitude, $\Delta \tilde{h}$ is the normalized



Squeeze Film Gas Lubrication, Fig. 6 2-D map of excursion amplitude of panel transducer of squeeze film guideway (Kobayashi et al. 2007, Ono et al. 2009)

excursion profile and is a function of the bearing plan-form coordinates, and the complex exponential notation is used to represent the simple harmonic time dependence at the driven frequency f_D , which would be the resonance frequency associated with $\Delta\tilde{h}$ in the resonance analysis of the transducer structure. The complete film thickness is

$$h = C + e\Delta h + a_{\text{excursion}}\Delta\tilde{h}e^{i(2\pi f_D)t} \quad (1)$$

where e is a load-induced bearing displacement and Δh is the normalized displacement profile that is dependent on the plan-form coordinates.

Governing Equations of Fluid Film

The isothermal gaseous lubrication theory with allowance of the squeeze film excursions is

$$\frac{\partial}{\partial t}(ph) + \frac{\partial}{\partial x}\left(\frac{U}{2}ph\right) = \frac{1}{12\mu}\text{div}(h^3p\text{grad } p) \quad (2)$$

Upon scaling

$$x, \text{grad}, t, h, p$$

respectively, with

$$l_x, L^{-1}, (2\pi f_D)^{-1}, C, p_a$$

Equation (2) assumes the dimensionless form

$$\frac{\partial\Psi}{\partial\tau} + \frac{\Lambda}{\sigma}\frac{\partial\Psi}{\partial\bar{x}} = \sigma^{-1}\nabla \cdot \left(\frac{H}{2}\nabla\Psi^2 - \Psi^2\nabla H\right) \quad (3)$$

With the presence of driven high-frequency squeeze oscillations, there is the adjunct condition of a steady-state squeeze film:

$$\int_{2\pi} \left[\Lambda \frac{\partial\Psi}{\partial\bar{x}} - \nabla \cdot \left(\frac{H}{2}\nabla\Psi^2 - \Psi^2\nabla H \right) \right] d\tau = 0 \quad (4)$$

Asymptotic Analysis of Squeeze Films

Asymptotic analysis of large- σ squeeze films presents the solution of (3) as the sum of two terms. The first term, Ψ_{∞} periphery, is derived by a formal limiting process of an unbounded σ

$$\frac{\partial\Psi_{\infty}}{\partial\tau} = 0 \quad (5)$$

It is not valid at the film periphery where $\Psi = H$ cannot be τ -independent. The second term is an edge solution, which describes a profile of Ψ_{edge} along an expanded normal coordinate, $\bar{\nu}\xi \perp$ periphery, such that

$$\bar{\nu} \cdot \nabla\Psi_{\text{edge}} = \sqrt{\sigma} \frac{\partial\Psi_{\text{edge}}}{\partial\xi_{\text{edge}}} \quad (6)$$

Accordingly, (3) rewritten for $\Psi_{\text{edge}}(\xi)$ yields

$$\frac{\partial\Psi_{\text{edge}}}{\partial\tau} = \frac{H_{\text{edge}}}{2} \frac{\partial^2\Psi_{\text{edge}}^2}{\partial\xi^2} \quad (7)$$

Ψ_{edge} is subject to the boundary conditions

$$\Psi_{\text{edge}}(\zeta = 0) = H_{\text{edge}}(\Psi_{\text{edge}} - \Psi_{\infty}, \partial\Psi_{\text{edge}}/\partial\zeta)_{\zeta \rightarrow \infty} = 0 \quad (8)$$

The steady-state condition of a squeeze film is applicable to (7), therefore

$$0 = \int_{2\pi} \frac{H_{\text{edge}}}{2} \frac{\partial^2 \Psi_{\text{edge}}^2}{\partial \zeta^2} d\tau \quad (9)$$

Integrate with respect to ζ from an unspecified lower limit to $\zeta \rightarrow \infty$, one finds

$$0 = \int_{2\pi} \frac{H_{\text{edge}}}{2} \frac{\partial \Psi_{\text{edge}}^2}{\partial \zeta} d\tau$$

Integrate again for the full range $0 \leq \zeta \rightarrow \infty$, one obtains

$$\Psi_{\infty}^2 = \frac{1}{\pi} \int_{2\pi} \frac{H_{\text{edge}}^3}{2} d\tau = 1 + \frac{3}{2} e_{\text{edge}}^2 \quad (10)$$

The steady-state condition of a squeeze film is also applicable to (4), with Ψ_{∞} substituted for Ψ , yielding

$$\Lambda \frac{\partial \Psi_{\infty}}{\partial \bar{x}} - \nabla \cdot \left(\frac{\bar{H}}{2} \nabla \Psi_{\infty}^2 - \Psi_{\infty}^2 \nabla \bar{H} \right) = 0 \quad (11)$$

Ψ_{∞} is solved from (11) with (10) furnishing the boundary conditions. The edge problem to solve for Ψ_{edge} requires numerical computation of Eq. (7).

Simple Circular Disk Squeeze Film

The simple circular disk concerns the problem treated by Salbu; referring to the equations above, the following special data values apply:

$$U = 0, \quad e = 0, \quad \Delta h = 1.$$

Due to rotational symmetry and a uniform film thickness, (11) is reduced to

$$-\frac{\partial}{\partial \zeta} \left(\frac{\zeta}{2} \frac{\partial \Psi_{\infty}^2}{\partial \zeta} \right) = 0 \quad (12)$$

Resulting in

$$\Psi_{\infty}^2 = 1 + \frac{3}{2} e_{\text{excursion}}^2 \quad (13)$$

The edge solution, as computed from (7), applies to the entire periphery of the squeeze film disk.

Axial load supported by the simple circular squeeze-film disk is calculated as

$$\begin{aligned} \frac{F_z}{\pi R^2 p_a} &= \int_{2\pi} \left[\int_0^1 \left(\frac{\Psi_{\infty}}{H} - 1 \right) \zeta d\zeta \right. \\ &\quad \left. + \int_0^{\infty} \frac{\Psi_{\text{edge}}}{\sqrt{\sigma} H_{\text{edge}}} d\zeta \right] \frac{d\tau}{\pi} \end{aligned} \quad (14)$$

Results for various values of $e_{\text{excursion}}$ are shown in Fig. 1.

Rectangular Squeeze Film Panel

A squeeze film panel to be used in a linear guideway would be studied in non-dimensional rectangular Cartesian coordinates that use

$$l_x, l_y, (2\pi f_D)^{-1}, C, p_a$$

to scale

$$x, y, t, h, p$$

respectively. Equation (3) can be rewritten as

$$\begin{aligned} \frac{\partial \Psi}{\partial \tau} + \frac{\Lambda_x}{\sigma_x} \frac{\partial \Psi}{\partial \bar{x}} &= \sigma_x^{-1} \left[\frac{\partial}{\partial \bar{x}} \left(\frac{H}{2} \frac{\partial \Psi^2}{\partial \bar{x}} - \Psi^2 \frac{\partial H}{\partial \bar{x}} \right) \right. \\ &\quad \left. + \alpha^{-2} \frac{\partial}{\partial \bar{y}} \left(\frac{H}{2} \frac{\partial \Psi^2}{\partial \bar{y}} - \Psi^2 \frac{\partial H}{\partial \bar{y}} \right) \right] \end{aligned} \quad (15)$$

For the asymptotic solution, (11) becomes

$$\begin{aligned} \frac{\partial}{\partial \bar{x}} \left(\frac{\bar{H}}{2} \frac{\partial \Psi_{\infty}^2}{\partial \bar{x}} - \Psi_{\infty}^2 \frac{\partial \bar{H}}{\partial \bar{x}} \right) &+ \alpha^{-2} \frac{\partial}{\partial \bar{y}} \left(\frac{\bar{H}}{2} \frac{\partial \Psi_{\infty}^2}{\partial \bar{y}} - \Psi_{\infty}^2 \frac{\partial \bar{H}}{\partial \bar{y}} \right) \\ &= \Lambda_x \frac{\partial \Psi_{\infty}}{\partial \bar{x}} \end{aligned} \quad (16)$$

The 2-D excursion map of the squeeze film, Fig. 6, depicts roughly diagonal nodes while mid-points of sides are peaks; those of horizontal sides are out of phase with those of vertical sides. Thus, adequate representations of the excursion profiles at the edges are

$$\begin{aligned} e_{\bar{x}=\mp \frac{1}{2}} &= e_{\text{excursion}} \cos \pi \bar{y} \\ e_{\bar{y}=\mp \frac{1}{2}} &= -e_{\text{excursion}} \cos \pi \bar{x} \end{aligned} \quad (17)$$

This furnishes boundary conditions for (16):

$$\begin{aligned} \Psi_{\infty}^2|_{\bar{x}=\mp \frac{1}{2}, \bar{y}} &= 1 + \frac{3}{2} e_{\text{excursion}}^2 \cos^2 \pi \bar{y} \\ \Psi_{\infty}^2|_{\bar{x}, \bar{y}=\mp \frac{1}{2}} &= 1 + \frac{3}{2} e_{\text{excursion}}^2 \cos^2 \pi \bar{x} \end{aligned} \quad (18)$$

According to (7), governing equations for the edge problems are

$$\begin{aligned} \frac{\partial \Psi_{\bar{x}=\mp \frac{1}{2}}}{\partial \tau} &= \frac{H_{\bar{x}=\mp \frac{1}{2}}}{2} \frac{\partial^2 \Psi_{\bar{x}=\mp \frac{1}{2}}^2}{\partial \zeta_{x^{\mp}}^2} \\ \frac{\partial \Psi_{\bar{y}=\mp \frac{1}{2}}}{\partial \tau} &= \frac{H_{\bar{y}=\mp \frac{1}{2}}}{2} \frac{\partial^2 \Psi_{\bar{y}=\mp \frac{1}{2}}^2}{\partial \zeta_{y^{\mp}}^2} \end{aligned} \quad (19)$$

The boundary conditions are

$$\begin{aligned} \Psi_{\bar{x}=\mp \frac{1}{2}}(\zeta_{x^{\mp}} = 0) &= H_{\bar{x}=\mp \frac{1}{2}} \\ \Psi_{\bar{y}=\mp \frac{1}{2}}(\zeta_{y^{\mp}} = 0) &= H_{\bar{y}=\mp \frac{1}{2}} \end{aligned} \quad (20)$$

And for $\zeta_{x^{\mp}}, \zeta_{y^{\mp}} \rightarrow \infty$

$$\begin{aligned} \left(\Psi_{\bar{x}=\mp\frac{1}{2}} - \Psi_{\infty}|_{\bar{x}=\mp\frac{1}{2}, \bar{y}}, \partial \Psi_{\bar{x}=\mp\frac{1}{2}} / \partial \xi_{x\mp} \right) &= 0 \\ \left(\Psi_{\bar{y}=\mp\frac{1}{2}} - \Psi_{\infty}|_{\bar{x}, \bar{y}=\mp\frac{1}{2}}, \partial \Psi_{\bar{y}=\mp\frac{1}{2}} / \partial \xi_{y\mp} \right) &= 0 \end{aligned} \quad (21)$$

Load supported by the rectangular squeeze film panel is calculated as the sum of an interior asymptotic solution and four boundary edge solutions. The interior asymptotic solution is

$$\frac{F_{\text{panel};\infty}}{l_x l_y p_a} = \int_{2\pi} \left\{ \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\Psi_{\infty}}{H} - 1 \right) d\bar{x} d\bar{y} \right\} \frac{d\tau}{2\pi} \quad (22)$$

The edge solution at either end of the panel length is

$$\frac{F_{\text{panel};\bar{x}=\mp\frac{1}{2}}}{l_x l_y p_a} = \int_{2\pi} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\int_0^{\infty} \frac{\Psi_{\bar{x}=\mp\frac{1}{2}} - \Psi_{\infty}}{H_{\bar{x}=\mp\frac{1}{2}}} d\xi_{x\pm} \right) \frac{d\bar{y}}{\sqrt{\sigma_x}} \frac{d\tau}{2\pi} \quad (23)$$

The edge solution at either end of the panel width is

$$\frac{F_{\text{panel};\bar{y}=\mp\frac{1}{2}}}{l_x l_y p_a} = \int_{2\pi} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\int_0^{\infty} \frac{\Psi_{\bar{y}=\mp\frac{1}{2}} - \Psi_{\infty}}{\alpha H_{\bar{y}=\mp\frac{1}{2}}} d\xi_{y\pm} \right) \frac{d\bar{x}}{\sqrt{\sigma_y}} \frac{d\tau}{2\pi} \quad (24)$$

Numerical Computations

Direct time-space computation of (3) can be performed as an initial value problem; e.g., to simulate the squeeze film start-up process, or dynamic simulation in conjunction with a supported object. In such treatments, (3) takes on the characteristics of a parabolic partial differential equation; hence, rules regarding numerical stability should be adhered to in the computation procedure. While such a computation is applied to large σ cases, the asymptotic behavior of the edge region that is of $O\{1/\sqrt{\sigma}\}$; grid spacing near the periphery should be selected to satisfy requirements of spatial resolution.

The condition of a steady-state squeeze film, (4), deals with a similar temporal integration. However, periodic closure is imposed in lieu of the appearance of an explicit τ -derivative. Although co-presence of Ψ and Ψ^2 precludes an exact computation, convergence by an iterative procedure should be readily achievable.

Key Applications

Flat top base of a metrology device

Guideway

Low-friction gimbal

Nomenclature

Roman letters	
C	Nominal clearance, m
e	Displacement amplitude, m
$e_{\text{excursion}}$	Excursion amplitude, m
f_D	Driven squeeze frequency, s^{-1}
F_z	Supported axial load, <i>Newton</i>
h	Film thickness, m
Δh	Normalized displacement profile
$\Delta \tilde{h}$	Normalized excursion profile
H	Normalized film thickness
\bar{H}	Squeeze period averaged H , $\equiv (2\pi)^{-1} \int_{2\pi} H d\tau$
\vec{i}	Base vector along sliding
l_x, l_y	Length and width of rectangular squeeze-film, m
L	Generic bearing dimension, m
p	Film pressure, <i>MPa</i>
p_a	Ambient pressure, <i>MPa</i>
P	Normalized film pressure, $\equiv p/p_a$
r	Radial coordinate, m
R	Radius of circular disk, m
t	Time, s
U	Surface sliding speed, m/s
x, y	Rectangular Cartesian coordinates, m
\bar{x}, \bar{y}	$= x/l_x, y/l_y$
Greek letters	
α	Aspect ratio of rectangular panel, $= l_y/l_x$
ϵ_{edge}	Normalized excursion amplitude at bearing edge, $= C^{-1} e_{\text{excursion}} \Delta \tilde{h}_{\text{edge}}$
$\vec{\phi}$	Film flux vector, $MPa \cdot m^2/s$
Λ	Generic bearing number, $= 6\mu UL(p_a C^2)^{-1}$
Λ_x	Guideway bearing number, $= 6\mu U l_x (p_a C^2)^{-1}$
μ	Gas viscosity, $Pa \cdot s$
\vec{v}	Inward directed unit normal at a periphery point of squeeze film.
σ	Squeeze number of a generic squeeze film, $\equiv 12\mu(2\pi f_D)(p_a C^2)^{-1} L^2$; squeeze number of circular disk, $\equiv 12\mu(2\pi f_D)(p_a C^2)^{-1} R^2$
σ_x	Squeeze number of rectangular panel, $\equiv 12\mu(2\pi f_D)(p_a C^2)^{-1} l_x^2$
τ	Normalized time, $\equiv 2\pi f_D t$

ζ	Edge coordinate, $\vec{v} \cdot \nabla \Psi_{\text{edge}} = \sqrt{\sigma} (\partial \Psi_{\text{edge}} / \partial \zeta)$
Ψ	$\equiv PH$
ζ	Non-dimensional radial coordinate, $\equiv r/R$
Mathematical symbols	
div	Divergence operator, m^{-1}
grad	Gradient operator, m^{-1}
∇	Non-dimensional gradient, $= L\text{grad}$

Cross-References

► [Mathematical Foundation of Fluid Lubrication Theory](#)

References

W.A. Gross, Gas lubrication. *Int. J. Sci. Technol.* **13**, 32 (1963)
H. Kobayashi, S. Yoshimoto, M. Miyatake, Float characteristics of a squeeze-film air bearing for a linear motion. *Tribol. Int.* **40**(3), 503–511 (2007)
Y. Ono, S. Yoshimoto, M. Miyatake, Impulse-load dynamics of squeeze film gas bearings for a linear motion guide. *J. Tribol.* **131**(4), 41706 (2009)
C.H.T. Pan, On asymptotic analysis of gaseous squeeze-film bearings. *J. Basic Eng.* **89**(3), 245–253 (1967)
C.H.T. Pan, The gaseous squeeze-film at moderately large squeeze numbers. *J. Basic Eng.* **92**(4), 766–781 (1970)
C.H.T. Pan, S.B. Malanoski, P.H. Broussard Jr., J.L. Burch, Theory and experiments of squeeze-film gas bearings, Part I-cylindrical journal bearing. *J. Basic Eng.* **88**(1), 191–198 (1966)
E.O.J. Salbu, Compressible squeeze films and squeeze bearings. *J. Basic Eng.* **86**(2), 355–364 (1964)
G.I. Taylor, P.G. Saffman, Effects of compressibility at low Reynolds numbers. *J. Aeronaut. Sci.* **24**, 553 (1957)
S. Yoshimoto, Y. Anno, Y. Sato, K. Hamanaka, Float characteristics of squeeze-film gas bearings with elastic hinges. *JSME Int. J. Ser. C* **40**(2), 353–359 (1997)

Squeezed Gas Film

► [Squeeze Film Gas Lubrication](#)

Squeeze-Film Flow

► [Fluid Inertia in Three-Dimensional Flows](#)

SSFE – Specific Surface Free Energy

► [Surface Free Energy](#)

ST – Surface Tension

► [Interfacial Energy](#)

Star Planetary

► [Epicyclic Gear Trains](#)

Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts

ILYA I. KUDISH
Department of Mathematics, Kettering University, Flint, MI, USA

Synonyms

[Starved heavily loaded EHL contacts](#)

Definition

Starvation effects correspond to reduced supply of lubricant to contact, which, in turn, leads to reduction of the lubrication film thickness and increase in frictional stresses. Under starvation this film thickness reduction can be very significant, even with very small reduction of lubricant supply. The significance of effects of starvation depends not only on supply of lubricant but also on the rheology of lubricating fluid and the slide-to-roll ratio. Starved lubrication conditions are typical for various high-speed gears and bearings.

Scientific Fundamentals

This topic here is the asymptotic analysis of steady isothermal heavily loaded line EHL contacts for two

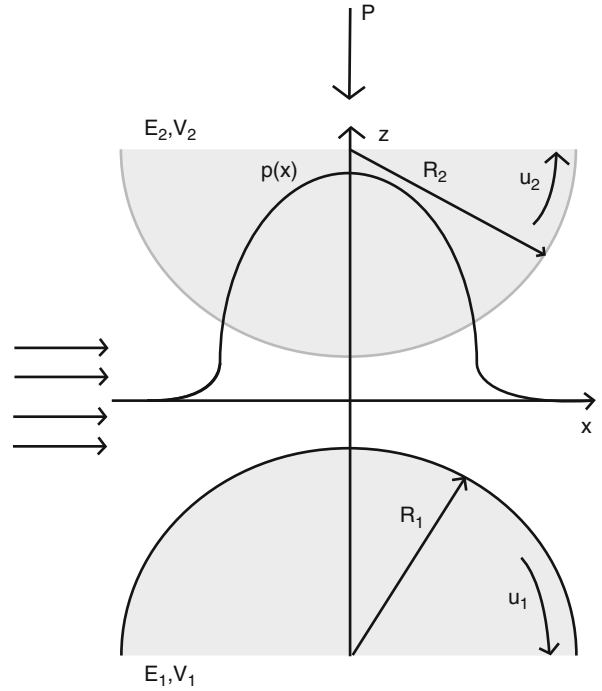
opposite limiting cases of pure rolling and relatively high slide-to-roll ratio. The problem is considered for two general classes of non-Newtonian lubricant rheologies when the shear strain and stress can be expressed as certain explicit functions of shear stress and strain, respectively. Some approximations of the generalized Reynolds equation for non-Newtonian fluids that resemble the Reynolds equation for a Newtonian fluid are obtained. The main idea of the method is analytical solution of the problem for the sliding frictional stress and the consequent reduction of the problem to asymptotic and numerical solutions of the problem for only the pressure and gap functions. The applied asymptotic procedure involves regular perturbations and an asymptotic approach similar to the one used in the case of Newtonian lubricants (see ► [Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts](#)). A number of examples illustrating application of the described technique are given. It is shown that in certain cases the asymptotic procedure described provides asymptotically correct solutions only for regimes of starved lubrication, while in other cases it is valid for both starved and fully flooded lubrication regimes.

Formulation of Isothermal EHL Problems for Non-Newtonian Fluids in Heavily Loaded Contacts

Consider a line-lubricated contact for infinite cylindrical solids of radii R_1 and R_2 made of elastic materials with Young's moduli E_1 and E_2 and Poisson's ratios ν_1 and ν_2 , respectively. The cylinders' surfaces are steadily moving, with linear velocities u_1 and u_2 , and pressed against each other with a normal force P . The lubricant separating the solids is an incompressible non-Newtonian fluid. Under classical assumptions (Kudish and Covitch 2010) of slow motion and narrow gap between the surfaces, the rheology of the lubricant can be described by the equations

$$\mu \frac{\partial u}{\partial z} = F(\tau) \text{ or } \Phi\left(\mu \frac{\partial u}{\partial z}\right), \quad (1)$$

where u is the lubricant velocity along the x -axis coinciding with the direction of fluid motion, $u=u(x, z)$; τ is the shear stress, $\tau=\tau(x, z)$; μ is the lubricant viscosity that depends on lubricant pressure p , $\mu=\mu(p)$; and F and Φ are given odd inverse to each other functions describing the lubricant rheology, $F(0)=\Phi(0)=0$. The coordinate system is introduced in such a manner that the x -axis is directed along the contact in the direction of motion, the y -axis is directed along cylinders' axes, and the z -axis is directed along the line connecting the cylinders' centers (Fig. 1).



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 1 General view of a lubricated contact

In the dimensionless variables for heavily loaded contact

$$\begin{aligned} \{x', a, c\} &= \frac{1}{a_H} \{x, x_i, x_e\}, p' = \frac{p}{p_H}, \{h', z'\} = \frac{1}{h_e} \{h, z\}, \\ \mu' &= \frac{\mu}{\mu_a}, \{\tau', f', \Phi'\} = \frac{\pi R'}{P} \{\tau, f, \Phi\}, F' = \frac{2h_e}{\mu_a(u_1 + u_2)} F, \end{aligned} \quad (2)$$

and parameters

$$s_0 = 2 \frac{u_2 - u_1}{u_2 + u_1}, V = \frac{24\mu_a(u_1 + u_2)R'^2}{a_H^3 p_H}, H_0 = \frac{2R'h_e}{a_H^2}, \quad (3)$$

the EHL problem can be reduced to the system of equations (Kudish and Covitch 2010) (further primes at the dimensionless variables are omitted)

$$\frac{d}{dx} \left\{ \frac{1}{\mu} \int_{-h/2}^{h/2} z F\left(f + H_0 z \frac{dp}{dx}\right) dz - h \right\} = 0, \quad (4)$$

$$\frac{1}{\mu} \int_{-h/2}^{h/2} F\left(f + H_0 z \frac{dp}{dx}\right) dz = s_0, \quad (5)$$

$$p(a) = p(c) = \frac{dp(c)}{dx} = 0, \quad (6)$$

$$H_0(h-1) = x^2 - c^2 + \frac{2}{\pi} \int_a^c p(t) \ln \left| \frac{c-t}{x-t} \right| dt, \quad (7)$$

$$\int_a^c p(t) dt = \frac{\pi}{2}. \quad (8)$$

In (2)–(8), function f is the sliding frictional stress, $f=f(x)$; h is the gap between the contact surfaces, $h=h(x)$; s_0 is the slide-to-roll ratio; a and c are the x -coordinates of the inlet and exit points of the contact; and H_0 is the dimensionless exit film thickness. In (2), p_H and a_H are the Hertzian maximum pressure and half-width of the contact, $p_H = \sqrt{\frac{E'P}{\pi R'}}$, $a_H = 2\sqrt{\frac{R'P}{\pi E'}}$, $\frac{1}{R'} = \frac{1}{R_1} \pm \frac{1}{R_2}$, $\frac{1}{E'} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}$; h_e is the dimensional lubrication film thickness at the exit from the contact; and μ_a is the lubricant viscosity at ambient pressure. In most cases, the rheology functions F and Φ and the lubricant viscosity μ also involve parameter V as well as some other parameters.

The system of equations (4)–(8) can be rewritten in the equivalent form (Kudish and Covitch 2010; Vorovich et al. 1974)

$$p(x) = R(x) \left[1 - \frac{1}{2\pi} \int_a^c \frac{dM(p, h)}{dt} \frac{dt}{R(t)(t-x)} \right], \quad (9)$$

$$R(x) = \sqrt{(x-a)(c-x)},$$

$$\int_a^c \frac{dM(p, h)}{dt} \frac{dt}{R(t)} = \pi(a+c), \quad \int_a^c \frac{dM(p, h)}{dt} \frac{tdt}{R(t)}$$

$$= \pi \left[\left(\frac{c-a}{2} \right)^2 + \frac{(a+c)^2}{2} - 1 \right], \quad (10)$$

$$M(p, h) = \frac{H_0}{\mu} \int_{-h/2}^{h/2} zF \left(f + H_0 z \frac{dp}{dx} \right) dz, \quad (11)$$

$$M(p, h) = H_0(h-1). \quad (12)$$

Therefore, for the given values of parameters a , V , and s_0 and other parameters involved, as well as for the given functions $\mu(p)$, $F(x)$, and $\Phi(x)$, the solution of the EHL problem is represented by parameters c , H_0 , and functions $f(x)$, $p(x)$, and $h(x)$.

Isothermal EHL Problem for Pure Rolling, Pre-critical Lubrication Regimes

One of the conditions most often encountered in practice is lubrication under almost purely rolling conditions. Consider the EHL problem for heavily loaded contact under pure rolling conditions. For the slide-to-roll ratio $s_0=0$, due to the fact that $F(x)$ is an odd function and $F(0)=0$, the solution of (5), is $f(x)=0$. In this case, the generalized Reynolds equation (5) is reduced to

$$\frac{d}{dx} \left\{ M \left(\mu, p, h, \frac{dp}{dx}, V, H_0 \right) - H_0 h \right\} = 0, \quad (13)$$

$$M(\mu, p, h, \frac{dp}{dx}, V, H_0) = \frac{H_0}{\mu} \int_{-h/2}^{h/2} zF \left(H_0 z \frac{dp}{dx} \right) dz. \quad (14)$$

Therefore, the lubrication problem is reduced to solution of the system of equations (13), (14), (6)–(8) or (9), (10), (12), and (14).

Heavily loaded EHL contact conditions are usually caused by low surface velocities and/or high load applied to the contact and/or a lubricant viscosity that experiences steep increase with pressure. In any of these cases, the EHL problem contains a small parameter $\omega \ll 1$. When the lubricant viscosity μ increases with pressure p moderately then the small parameter $\omega = V \ll 1$. In the cases when heavy loading is caused by fast increase of the lubricant viscosity μ with pressure p from 1 to $\mu(1) \gg 1$, the small parameter can be taken as $\omega = 1/\ln(\mu(1)) \ll 1$. In particular, for exponential law $\mu = \exp(Qp)$ (where $Q = \alpha_p p_H$, α_p is the pressure viscosity coefficient) the latter definition leads to $\omega = Q^{-1} \ll 1$.

Assuming that ε_q is the characteristic size of the inlet zone, pre-critical regimes can be defined as such regimes for which $\mu(\varepsilon_q^{1/2}) = O(1)$, $\omega \ll 1$. A detailed classification and analysis of over-critical regimes is given in Kudish and Covitch (2010). To consider pre-critical lubrication regimes it is necessary to make an assumption about the rheological function

$$F(H_0 \varepsilon_q^{-1/2} y(t)) = V^{-k} (\omega^{-l} \varepsilon_q^{-1/2} H_0^{n+1})^{1/m} F_0(y(t)) + \dots,$$

$$F_0(y(t)) = O(1) \quad \text{for } y(t) = O(1), \quad (15)$$

where k , l , m , and n are constants, $m > 0$. Using this assumption and following the asymptotic technique described in ► [Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts](#) and Kudish and Covitch (2010) one obtains systems of asymptotic equations for $q(r)$ and A in the inlet zone

$$q(r) = \sqrt{2r} \left[1 - \frac{1}{2\pi} \int_0^\infty \frac{d}{dt} M_0(q, h_q, \mu_q, t) \frac{dt}{\sqrt{2t(t-r)}} \right],$$

$$\pi\alpha_1 = \int_0^\infty \frac{d}{dt} M_0(q, h_q, \mu_q, t) \frac{dt}{\sqrt{2t}}, \quad (16)$$

$$M_0(p, h, \mu, x) = \frac{A^{\frac{m+n+1}{m}}}{\mu} \int_{-h/2}^{h/2} z F_0 \left(z \frac{dp}{dx} \right), \quad (17)$$

$$A(h_q - 1) = M_0(q, h_q, \mu_q, r), \quad (18)$$

for fully flooded lubrication regimes and

$$h_q = 1 \quad (19)$$

for starved lubrication regimes as well as asymptotic equations for $g(s)$ and β_1 in the exit zone

$$g(s) = \sqrt{-2s} \left[1 - \frac{1}{2\pi} \int_{-\infty}^0 \frac{d}{dt} M_0(g, h_g, \mu_g, t) \frac{dt}{\sqrt{2t(t-s)}} \right],$$

$$\beta_1 = \frac{1}{\pi} \int_{-\infty}^0 \frac{d}{dt} M_0(g, h_g, \mu_g, t) \frac{dt}{\sqrt{-2t}}, \quad (20)$$

$$A(h_g - 1) = M_0(g, h_g, \mu_g, s), \quad (21)$$

for fully flooded lubrication regimes and

$$h_g = 1, \quad (22)$$

for starved lubrication regimes. As a byproduct of this analysis, a formula for the lubrication film thickness is obtained as follows:

$$H_0 = A \left(V^{km} \omega^l \varepsilon_q^{\frac{3m+1}{2}} \right)^{\frac{1}{m+n+1}} + \dots, \quad A(\alpha_1) = O(1), \omega \ll 1, \quad (23)$$

where $A=A(\alpha_1)$ is an unknown non-negative constant independent from ω and ε_q , which is determined by the solution of the problem in the inlet zone; $\alpha_1 < 0$ and $\beta_1 > 0$ are given and unknown constants in the expressions for $a = -1 + \alpha_1 \varepsilon_q$, $\alpha_1 = O(1)$, and $c = 1 + \beta_1 \varepsilon_q$, $\beta_1 = O(1)$, respectively; and functions $h_q(r)$, $h_g(s)$, $\mu_q(q)$, and $\mu_g(g)$ are the main terms of asymptotic expansions of h and μ in the inlet and exit zones. Functions $q(r)$ and $g(s)$ are determined by the relationships $p(x) = \varepsilon_q^{1/2} q(r) + \dots$, $r = (x - a)/\varepsilon_q = O(1)$, and $p(x) = \varepsilon_q^{1/2} g(s) + \dots$, $s = (x - c)/\varepsilon_q = O(1)$, in the inlet and exit zones, respectively.

Starved lubrication regimes are determined by small supply of lubricant at the inlet in the contact. That leads to a nearly constant gap within the entire lubricated contact. Mathematically it is translated in smallness of the ratio $\varepsilon_q^{3/2}/H_0 \ll 1$, which requires that $\varepsilon_q \ll \omega^{\frac{2l}{3m+2}}$, $\omega \ll 1$, and the expressions for the gap in the inlet and exit zones become $h_q(r)=1$, $h_g(s)=1$ (see (19) and (22)). For fully flooded lubrication regimes determined by the relationship $H_0 = O(\varepsilon_q^{3/2})$, $\omega \ll 1$, one obtains $\varepsilon_q = \omega^{\frac{2l}{3m+2}}$, $\omega \ll 1$, while the gap functions $h_q(r)$ and $h_g(s)$ are determined by (18) and (21), respectively. For fully flooded lubrication regimes

$$H_0 = A(V^{km} \omega^l)^{\frac{3}{3m+2}} + \dots, \quad A(\alpha_1) = O(1), \omega \ll 1. \quad (24)$$

Consider some examples of pre-critical lubrication regimes.

Example 1. For a Newtonian fluid (see (1))

$$F(x) = \frac{12H_0}{V} x, \quad \Phi(x) = \frac{V}{12H_0} x, \quad M = \frac{H_0^3}{V} \frac{h^3}{\mu} \frac{dp}{dx}. \quad (25)$$

In (46) parameters $k=0, l=m=n=1$ or $k=1, l=0, n=1$ for $\omega=V \ll 1$ and $k=1, l=0, m=n=1$ for $\omega=Q^{-1} \ll 1$. Therefore, (23) and (24) are reduced to

$$H_0 = A_s (V \varepsilon_q^2)^{1/3}, \quad A_s = O(1), \varepsilon_q \ll \varepsilon_f = V^{2/5}, \omega \ll 1, \quad (26)$$

for starved lubrication regimes and for fully flooded lubrication regimes

$$H_0 = A_f V^{3/5}, \quad A_f = O(1), \varepsilon_q = O(V^{2/5}), \omega \ll 1. \quad (27)$$

Example 2. For a fluid with the Ostwald-de Waele (power law) rheology

$$F(x) = \frac{12H_0}{V_n} |x|^{(1-n)/n} x, \quad V_n = V \left(\frac{P}{\pi R' G} \right)^{(n-1)/n}, \quad (28)$$

where G is the characteristic shear modulus of the fluid. In this case (14) reads

$$M = \frac{24n}{2n+1} \frac{H_0^{\frac{2n+1}{n}}}{V_n} \frac{1}{\mu} \left(\frac{h}{2} \right)^{\frac{2n+1}{n}} \left| \frac{dp}{dx} \right|^{\frac{1-n}{n}} \frac{dp}{dx}. \quad (29)$$

while parameters in (46) are $k=0, l=m=n$ for $\omega=V_n \ll 1$ and $k=1, l=0, m=n$ for $\omega=Q^{-1} \ll 1$. Therefore, for starved lubrication regimes

$$H_0 = A_s (V_n^{\frac{3n+1}{n}} \varepsilon_q^{\frac{2}{n}})^{\frac{1}{2n+1}}, \quad A_s = O(1), \quad \varepsilon_q \ll \varepsilon_f = V_n^{\frac{2n}{3n+2}}, \quad \omega \ll 1, \quad (30)$$

and for fully flooded lubrication regimes

$$H_0 = A_f V_n^{\frac{3n}{3n+2}}, \quad A_f = O(1), \quad \varepsilon_q = O(\varepsilon_f), \quad \omega \ll 1. \quad (31)$$

For $n=1$ the Ostwald-de Waele rheology is reduced to the Newtonian one, and (29) for M is reduced to (25) for a lubricant with Newtonian rheology.

Example 3. For a lubricant with Reiner-Philippoff-Carreau rheology

$$\Phi(x) = \frac{V}{12H_0} x \left\{ \eta + (1 - \eta) \left[1 + \left(\frac{V}{12H_0 G_0} |x| \right)^m \right]^{\frac{n-1}{m}} \right\},$$

$$\eta = \frac{\mu_\infty}{\mu_0}, \quad G_0 = \frac{\pi R' G}{P}, \quad (32)$$

where μ_0 and μ_∞ are the lubricant viscosities at the shear rates $\partial u / \partial z = 0$ and $\partial u / \partial z = \infty$, respectively, which are scaled according to formulas (2), and G is the characteristic shear stress of the fluid, $G > 0$, n and m are constants, $0 \leq n \leq 1$, $m > 0$.

There are two limiting cases for (32). If shear stress τ is relatively small then the lubricant behaves like a Newtonian fluid, that is,

$$\Phi(x) = \frac{V}{12H_0} x \quad \text{if} \quad \frac{V}{12H_0 G_0} \mu \frac{\partial u}{\partial z} \ll 1. \quad (33)$$

The Reynolds equation coincides with the one for Newtonian fluid. In case of relatively large shear stress τ , the lubricant behaves according to power law

$$\Phi(x) = (1 - \eta) \left(\frac{V}{12H_0} \right)^n \left(\frac{|x|}{G_0} \right)^{n-1} x \quad \text{if} \quad \frac{V}{12H_0 G_0} \mu \frac{\partial u}{\partial z} \gg 1. \quad (34)$$

Solving (34) for x provides the expression for $F(x)$ in the form similar to (28)

$$F(x) = \frac{12H_0}{V(1 - \eta)^{1/n}} \left[\frac{|x|}{G_0} \right]^{(1-n)/n} x, \quad \text{if} \quad \frac{V}{12H_0 G_0} \mu \frac{\partial u}{\partial z} \gg 1. \quad (35)$$

It this case, (14) reads

$$M = \frac{24n}{2n+1} \frac{H_0^{\frac{2n+1}{n}}}{V(1 - \eta)^{1/n} G_0^{(1-n)/n}} \frac{1}{\mu} \left(\frac{h}{2} \right)^{\frac{2n+1}{n}} \frac{dp}{|dx|} \frac{dp}{dx}. \quad (36)$$

There are three distinct cases. In both cases when $\frac{V}{12H_0 G_0} \left| \mu \frac{\partial u}{\partial z} \right| \ll 1$ or $\frac{V}{12H_0 G_0} \left| \mu \frac{\partial u}{\partial z} \right| = O(1)$ for $\omega \ll 1$, the formulas for the film thickness H_0 coincide with the ones for a Newtonian fluid (see Example 1). However, the coefficients of proportionality A_s and A_f in the latter case

differ from the Newtonian case and depend on the values of parameters n , m and G_0 . In the case when $V / (12H_0 G_0) \left| \mu \partial u / \partial z \right| \gg 1$ for $\omega \ll 1$ (here the small parameter ω and the corresponding powers k , l , m , and n are determined as in Example 2), the rheological function F is well approximated by the power law represented by formulas (29) in which parameter V_n must be replaced by $V_n = V [G_0(1 - \eta)]^{\frac{1}{n}} / G_0 \ll 1$ (which becomes obvious from comparison formulas (29) and (34)). In this case, it is easy to determine function

$$F(x) = \frac{1}{(1 - \eta)^{1/n}} \frac{12H_0}{V} \left\{ \frac{|x|}{G_0} \right\}^{\frac{1-n}{n}} x.$$

Therefore, the formulas for film thickness H_0 coincide with the ones for a fluid with the power law rheology (see Example 2). The above conditions on $\frac{V}{12H_0 G_0} \left| \mu \frac{\partial u}{\partial z} \right| \gg 1$ impose some limitations on the problem parameters that are not restrictive in the limiting case of Newtonian behavior and are fairly restrictive in the limiting case of “power law” behavior (Kudish and Covitch 2010). Practically, these restrictions mean that the pure rolling operating conditions for which a fluid with Reiner-Philippoff-Carreau rheology demonstrates non-Newtonian behavior hardly exist.

Isothermal EHL Problem for Relatively Large Sliding, Pre-critical Lubrication Regimes

Consider the EHL problems with non-Newtonian lubricants in cases of heavily loaded contacts with relatively large slide-to-roll ratios s_0 . It is necessary to introduce a small function

$$v(x) = \frac{H_0 h dp}{2f dx} \ll 1, \quad \omega \ll 1, \quad (37)$$

representing the ratio of the rolling $0.5H_0 h dp/dx$ and sliding f frictional stresses. Assume that this function is small in the entire contact or just in the inlet zone (see below) of the contact, that is, $v(x) \ll 1$, $\omega \ll 1$, regardless of the particular definition of parameter ω .

For $v(x) \ll 1$ the representation for $f(x)$ is searched in the form

$$f(x) = f_0(x) + v(x)f_1(x) + v^2(x)f_2(x) + v^3(x)f_3(x) + O(v^4(x)), \quad (38)$$

where functions $f_0(x)$, $f_1(x)$, $f_2(x)$, and $f_3(x)$ are the consecutive terms of the asymptotic of $f(x)$ that have to be determined. Originally, this technique was introduced in Kudish (1978, 1979, 1982, 1983).

Substituting representation (38) for $f(x)$ into (5) and expanding it for $v(x) \ll 1$ leads to

$$\begin{aligned} f_0(x) &= \Phi\left(\frac{\mu s_0}{h}\right), \quad f_1(x) = 0, \quad f_2(x) = -\frac{f_0^2 F''(f_0)}{6F'(f_0)}, \\ f_3(x) &= 0, \quad \dots \end{aligned} \quad (39)$$

The obtained solution for $f(x)$ allows for its elimination from the set of the problem unknowns and reduction of the problem to determining just two functions: pressure p and gap h . Substituting (37)–(39) into (4) results in the approximate generalized Reynolds equation

$$\frac{d}{dx} \left\{ M \left(\mu, p, h, \frac{dp}{dx}, V, s_0, H_0 \right) - H_0 h \right\} = 0. \quad (40)$$

where the expressions for function M depend on the number of retained terms in (38) and (39). If the first one or two terms are retained in the expansion for $f(x)$ from (38) (i.e., $f(x) = f_0(x) + O(v^2(x))$, $v(x) \ll 1$) then

$$M = \frac{H_0^2 h^3 F'(f_0)}{12\mu} \frac{dp}{dx}, \quad (41)$$

while if the first three or four terms are retained in the expansion for $f(x)$ from (38) (that is, $f(x) = f_0(x) + v(x)f_2(x) + O(v^4(x))$, $v(x) \ll 1$) then

$$M = \frac{H_0^2 h^3 F'(f_0)}{12\mu} \frac{dp}{dx} \left\{ 1 + \frac{H_0^2 h^2}{8} \left[\frac{F''(f_0)}{5F'(f_0)} - \frac{1}{3} \left(\frac{F''(f_0)}{F'(f_0)} \right)^2 \right] \left(\frac{dp}{dx} \right)^2 \right\}. \quad (42)$$

In cases when the lubricant rheology is given by the second equation in (1), the derivatives of function $F(f)$ with respect to f involved in (41) and (42) can be expressed in the following way:

$$\begin{aligned} F'(f_0) &= \frac{1}{\Phi'(\lambda)}, \quad F''(f_0) = -\frac{\Phi''(\lambda)}{[\Phi'(\lambda)]^3}, \\ F'''(f_0) &= \frac{3[\Phi''(\lambda)]^2 - \Phi'''(\lambda)\Phi'(\lambda)}{[\Phi'(\lambda)]^5}, \quad \lambda = \frac{\mu s_0}{h}, \end{aligned} \quad (43)$$

where function $\Phi(\lambda)$ is differentiated with respect to λ . Therefore, in these cases the expressions for function M that correspond to (41) and (42) can be represented in the forms

$$M = \frac{H_0^2 h^3}{12\mu \Phi'(\lambda)} \frac{dp}{dx}, \quad \lambda = \frac{\mu s_0}{h} \quad (44)$$

$$\begin{aligned} M &= \frac{H_0^2 h^3}{12\mu \Phi'(\lambda)} \frac{dp}{dx} \left\{ 1 + \frac{H_0^2 h^2}{120} \frac{4[\Phi''(\lambda)]^2 - 3\Phi'''(\lambda)\Phi'(\lambda)}{[\Phi'(\lambda)]^4} \left(\frac{dp}{dx} \right)^2 \right\}, \\ \lambda &= \frac{\mu s_0}{h}, \end{aligned} \quad (45)$$

respectively.

As a result of this analysis for the regimes for which $v(x) \ll 1$ for $a \leq x \leq c$, the EHL problem can be reduced to a system of equations (9), (10), and (12) where function M is determined by one of the formulas (41), (42), (44), or (45).

To proceed with the asymptotic analysis of the pre-critical lubrication regimes it is necessary to make an assumption about function Φ

$$\begin{aligned} \Phi\left(\frac{\mu s_0}{h}\right) &= V^k \omega^l |s_0|^m \text{sign}(s_0) H_0^{-n} \Phi_0\left(\frac{\mu}{h}\right), \\ \Phi_0\left(\frac{\mu}{h}\right) &= O(1), \quad r = O(1), \quad \omega \ll 1, \end{aligned} \quad (46)$$

where $\Phi_0(r)$ is a certain function of r while l , m , k , and n are certain constants. To estimate functions involved in (38) and (39), one obtains the asymptotic behavior of Φ' , Φ'' , and Φ''' in the inlet zone as follows:

$$\begin{aligned} \Phi'\left(\frac{\mu s_0}{h}\right) &= V^k \omega^l |s_0|^{m-1} H_0^{-n} \Phi_1\left(\frac{\mu}{h}\right), \quad \Phi_1\left(\frac{\mu}{h}\right) = O(1), \\ r &= O(1), \end{aligned} \quad (47)$$

$$\begin{aligned} \Phi''\left(\frac{\mu s_0}{h}\right) &= V^k \omega^l |s_0|^{m-2} H_0^{-n} \Phi_2\left(\frac{\mu}{h}\right), \quad \Phi_2\left(\frac{\mu}{h}\right) = O(1), \\ r &= O(1), \end{aligned} \quad (48)$$

$$\begin{aligned} \Phi'''\left(\frac{\mu s_0}{h}\right) &= V^k \omega^l |s_0|^{m-3} H_0^{-n} \Phi_3\left(\frac{\mu}{h}\right), \quad \Phi_3\left(\frac{\mu}{h}\right) = O(1), \\ r &= O(1), \end{aligned} \quad (49)$$

where $\Phi_1(\mu/h)$, $\Phi_2(\mu/h)$, and $\Phi_3(\mu/h)$ are certain functions of r . Similar relationships for Φ , Φ' , Φ'' , and Φ''' hold in the exit zone. Considering orders of magnitude of the terms of the generalized Reynolds equation in the inlet zone leads to a formula for the film thickness

$$H_0 = A(V^k \omega^l |s_0|^{m-1} \varepsilon_q^2)^{\frac{1}{m+2}}, \quad A = O(1), \quad \omega \ll 1. \quad (50)$$

In the latter equation, A is a coefficient of proportionality that depends only on the specifics of the rheology function F (and/or Φ) and the lubricant viscosity μ and it is independent of ω , s_0 , and ε_q . The value of coefficient A can be obtained experimentally or by numerical solution of the system of asymptotically valid in the inlet zone (16) and (18) with the corresponding function M_0 . Equation (50) is valid for both starved and fully flooded lubrication regimes. For fully flooded lubrication regimes

$$H_0 = A_f (V^k \omega^l |s_0|^{m-1})^{\frac{3}{3n+2}}, \quad A_f = O(1), \quad \varepsilon_q = O(\varepsilon_f),$$

$$\varepsilon_f = (V^k \omega^l |s_0|^{m-1})^{\frac{2}{3n+2}}, \quad \omega \ll 1. \quad (51)$$

The value of ε_f (see (51)) represents the characteristic size of the inlet zone in the case of fully flooded pre-critical lubrication regimes. In (51), coefficient A_f is independent of ω , s_0 , V , ε_0 , and ε_q and it can be determined experimentally or numerically.

The conditions under which the applied perturbation analysis for pre-critical regimes is valid is equivalent to the validity of the estimate $v(x) \ll 1$, $\omega \ll 1$ in the inlet zone. It can be shown that it is valid if

$$\varepsilon_q \ll \varepsilon_v = (V^k \omega^l |s_0|^{m+n+1})^{\frac{2}{3n+2}}, \quad \omega \ll 1. \quad (52)$$

By comparing the magnitudes of the values of ε_f and ε_v , it can be established that the above analysis is valid for both starved and fully flooded regimes if $\varepsilon_f = O(\varepsilon_v)$, $\omega \ll 1$, which for $n+2 > 0$ is equivalent to the estimates

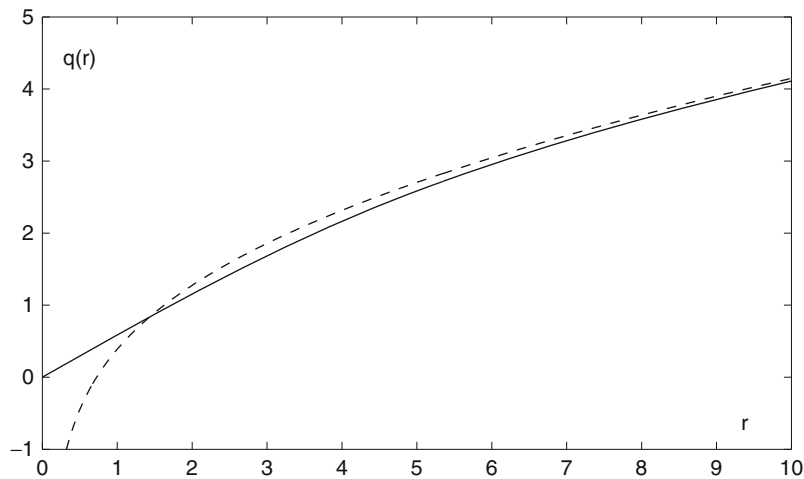
$$|s_0| \gg 1 \text{ or } |s_0| = O(1), \quad \omega \ll 1. \quad (53)$$

Numerical Solutions of Asymptotic Equations for Newtonian Fluids in Pre-critical Lubrication Regimes

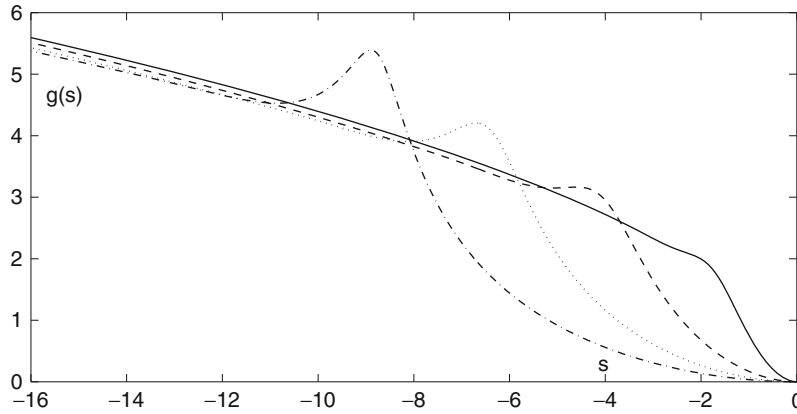
It is assumed that the lubricant viscosity satisfies the exponential law $\mu = e^{Qp}$. Therefore, for $\varepsilon_q = V^{2/5}$ in the inlet and exit zones the viscosity is $\mu_q = \exp(Q_{0q})$ and

$\mu_g = \exp(Q_{0g})$, $Q_0 = QV^{1/5}$. To obtain numerical solutions, the methods described in ► [Numerical Stability and Precision in Elastohydrodynamic Lubrication \(EHL\)](#) and in Kudish and Covitch (2010) are used. They demonstrated convergence and stability in the inlet zone for starved and fully flooded lubrication regimes. In the exit zone they also converge well for relatively small values of A and Q_0 without regularization, for other values of A and Q_0 the regularization is used that is described in detail in Kudish and Covitch (2010).

Consider some examples for Newtonian fluids. First, regimes of starved lubrication for which $\varepsilon_q^{1/2} = Q^{-1} \ll V^{1/5}$, $Q \gg 1$ and, therefore, $\mu_q(q) = e^q$ and $\mu_g(g) = e^g$ (that is, $Q_0 = 1$) are analyzed. The absolute error of calculations is chosen to be not higher than $\varepsilon = 10^{-4}$. To check the convergence of the numerical scheme for $A=2$, three series of calculations were done for $N=400$, $\Delta r=0.0625$, $N=800$, $\Delta r=0.03125$, and $N=1,600$, $\Delta r=0.015625$ ($N\Delta r=25$). The solution precision was reached after 8–13 iterations. The maximum relative errors of the solutions obtained for $N=400$, $N=800$, and $N=1,600$ in the values of α_1 , β_1 , $q(r)$, and $g(s)$ were found to be not greater than 0.58% and 0.33%. Therefore, the rest of calculations were done for $N=800$, $\Delta r=0.03125$. The graphs of $q(r)$ for $A=2$ and $g(s)$ for four values of the parameter $A=1, 2, 3$, and 4 are given in Figs. 2 and 3. For these values of A the corresponding values of α_1 are equal to -0.5287 , -1.4515 , -2.5887 , and -3.8782 while the values of β_1 are equal to 0.3646, 0.8966, 1.5017, and 2.1552. For different values of A the curves of $q(r)$



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 2 Main term of the asymptotic of the pressure distribution $q(r)$ (solid curve) and the asymptote of the Hertzian pressure $q_a(r) = \sqrt{2r} + \frac{\alpha_1}{\sqrt{2r}}$ (dashed curve) in the inlet zone of a starved lubricated contact for $Q_0=1$ and $A=2$



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 3 Main term of the asymptotic of the pressure distribution $g(s)$ in the exit zone of a starved lubricated contact for $Q_0=1$, $A=1$ (solid curve), $A=2$ (dashed curve), $A=3$ (dotted curve), and $A=4$ (dash-dotted curve)

resemble each other. Therefore, just one curve of $q(r)$ (solid curve) and for comparison the curve of the Hertzian pressure asymptote $q_a(r) = \sqrt{2r} + \alpha_1/\sqrt{2r}$ (dashed curve) for $A=2$ are given in Fig. 2. It can be seen from Fig. 2 that $q(r)$ is a monotonically increasing function of r that approaches its asymptote $q_a(r)$ and does not exhibit any signs of instability or oscillations. Figure 3 demonstrates the behavior of all four curves of $g(s)$ (solid, dashed, dotted, and dash-dotted curves correspond to $A=1, 2, 3$, and 4 , respectively). Each of the curves of $g(s)$ possesses a local maximum (pressure spike) that shifts closer to the center of the contact (to $s=-\infty$) and increases in value as the values of $|\alpha_1|$ and A increase. As in the case of $q(r)$ for large s the behavior of $g(s)$ practically coincides with the behavior of the asymptote of the Hertzian pressure $g_a(s) = \sqrt{-2s} - \beta_1/\sqrt{-2s}$ and does not exhibit any oscillations or signs of instability.

For the case of constant viscosity in the inlet and exit zones, $\mu_q(q)=\mu_g(g)=1$ and $Q_0=0$. Due to the fact that in this case solutions in the inlet and exit zones approach their asymptotes slower than for the case of exponential viscosity, it is chosen to do calculations for $N=1,600$ and $\Delta r=0.03125$. For $Q_0=0$ in the inlet and exit zones for $A=1, 2, 3$, and 4 one obtains the values of α_1 to be equal to -0.4999 , -1.4315 , -2.6279 , and -4.0345 while the values of β_1 are equal to 0.5080 , 1.4363 , 2.6561 , and 4.1274 , respectively. The numerical solutions for $q(r)$ in the inlet zone qualitatively resemble the one in Fig. 2. From the data presented above it follows that for the same values of constant A the values of α_1 for the cases of $\mu_q(q)=1$ and $\mu_q(q)=e^q$ do not vary much. Effectively, the above behavior indicates that for the

same values of ε_q and A the inlet zone is wider in the cases of constant viscosity ($\mu_q(q)=1$) compared with the cases when $\mu_q(q)=e^q$.

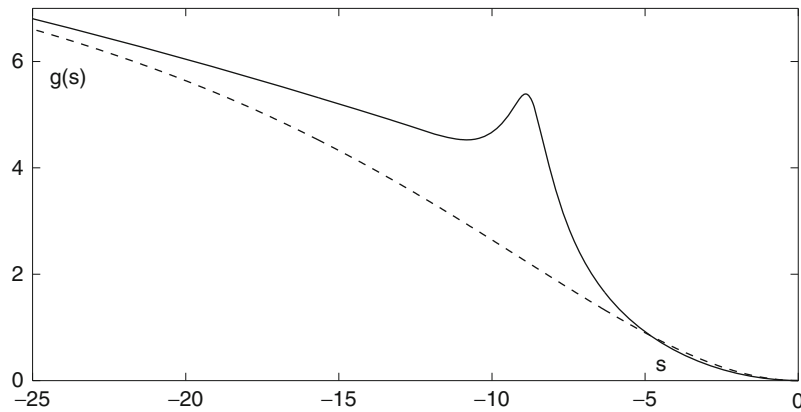
For starved lubrication regimes and constant viscosity $\mu_g(g)=1$ in the exit zone the solutions for $g(s)$ are monotonically decreasing functions of s that approach $\sqrt{-2s}$ as $s \rightarrow -\infty$. Also, it happens slower than in the corresponding cases of $\mu_g(g)=e^g$. Therefore, for $\varepsilon_q^{1/2} \ll Q^{-1} \ll V^{1/5}$, $Q \gg 1$ and $\mu_g(g)=1$ the solution functions $g(s)$ differ significantly from the solutions for $g(s)$ obtained for $\varepsilon_q^{1/2} = Q^{-1} \ll V^{1/5}$, $Q \gg 1$, $\mu_g(g)=e^g$, and depicted in Fig. 3. That can be seen from Fig. 4, in which for $A=4$ graphs of two functions $g(s)$ for $\mu_g(g)=1$ (solid curve) and $\mu_g(g)=e^g$ (dashed curve) are presented. Moreover, from the presented data it is clear that for regimes of starved lubrication for the same values of ε_q and constant A the exit zone for constant viscosity is also larger than for exponential viscosity ($\mu_g(g)=e^g$) and the difference in size increases as A increases.

For starved lubrication regimes and $Q_0=0$ a simple analytical property of the solution and numerical results provide for the formula for the film thickness (Kudish and Covitch 2010)

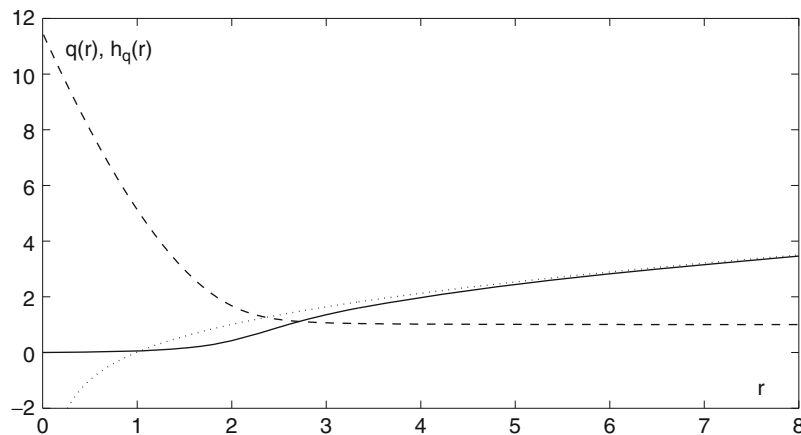
$$H_0 = 1.571(a+1)^{2/3} V^{1/3}, \quad (54)$$

where a is the inlet coordinate.

Now, consider some results for pre-critical fully flooded lubrication regimes for exponential viscosity $\mu(p) = \exp(Qp)$ and $\mu_q(q) = \exp(Q_0q)$, $\mu_g(g) = \exp(Q_0g)$, and $Q_0 = QV^{1/5} = O(1)$ for $\varepsilon_q = V^{2/5} = O(Q^{-2})$, $Q \gg 1$. For fully flooded lubrication regimes in the inlet zone the iterations



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 4 Main term of the asymptotic of the pressure distribution $g(s)$ in the exit zone of a starved lubricated contact for $A=4$ and viscosity $\mu_g(g)=1$ (solid curve) and $\mu_g(g)=e^g$ (dashed curve)



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 5 Main terms of the asymptotic distributions of pressure $q(r)$ (solid curve), gap $h_q(r)$ (dashed curve), and Hertzian pressure asymptote $q_a(r) = \sqrt{2r} + \frac{\alpha_1}{\sqrt{2r}}$ (dotted curve) in the inlet zone of a fully flooded lubricated contact for $A=0.525$ and $Q_0=1$

also converge to stable solutions, however, it takes more iterations. For example, for $Q_0=1$, $A=0.525$, $N=480$, $\Delta r=0.03125$, and the absolute precisions $\varepsilon=0.01$, $\varepsilon=0.001$, and $\varepsilon=0.0001$ the solutions in the inlet zone converged after 155, 332, and 517 iterations, respectively. The solutions obtained for these precision levels are as follows: $\alpha_1=-1.8853$, $h_q(\Delta r/2)=10.5904$, $\alpha_1=-1.9715$, $h_q(\Delta r/2)=11.2597$, and $\alpha_1=-1.9806$, $h_q(\Delta r/2)=11.3310$, respectively. For $Q_0=1$, $A=0.525$, $N=800$, and $\Delta r=0.015625$ the graphs of $q(r)$, $h_q(r)$, and the asymptote $q_a(r)$ of the Hertzian pressure are given in Fig. 5.

For $Q_0=1$ and several values of coefficient A the values of the inlet coordinate α_1 and gap $h_q(\Delta r/2)$ are presented in Table 1. This data gives an idea of how quickly lubricant starvation develops (i.e., the proximity of the inlet coordinate α_1 to zero) and how it affects the lubrication film thickness H_0 , which is directly proportional to A . It is not unexpected that when coefficient A approaches its limiting value the gap at the inlet point increases (without bound). Note that $A=0.525$ is relatively close to its limiting value, which requires many iterations. The evidence of that is in the fact that for $Q_0=1$, $A=0.535$, $N=480$, $\Delta r=0.03125$, and $\varepsilon=0.001$ one obtains $\alpha_1=-4.3238$ and

$h_q(\Delta r/2) = 32.9620$, that is, the value of $h_q(\Delta r/2)$ almost tripled while the value of A is increased from 0.525 to 0.535 by just 0.01 (or 1.9%). Also, the data from Table 1 shows that the size of the entire inlet zone is usually small (i.e., for $V \ll 1$ it is about $6V^{2/5} \ll 1$). The numerical analysis of the fully flooded inlet zone and formula (51) for $|a| \gg 1$, $a < 0$ leads to some formulas for the lubrication film thickness H_0

$$H_0 = AV^{3/5}, \quad A = 0.535, \quad Q_0 = 1; \quad A = 0.676, \quad Q_0 = 2. \quad (55)$$

In the exit zone, for fully flooded pre-critical regimes for $Q_0 = 1$, $A = 0.525$, $N = 800$, $\Delta r = 0.00390625$, and $\varepsilon = 0.0001$, the solution converges after 17 iterations to $\beta_1 = 0.120$ and $\min h_g(s) = 0.773$. For pre-critical fully flooded lubrication regimes in the exit zone for $Q_0 \leq 2.5$, the pressure distribution $g(s)$ behaves monotonically with

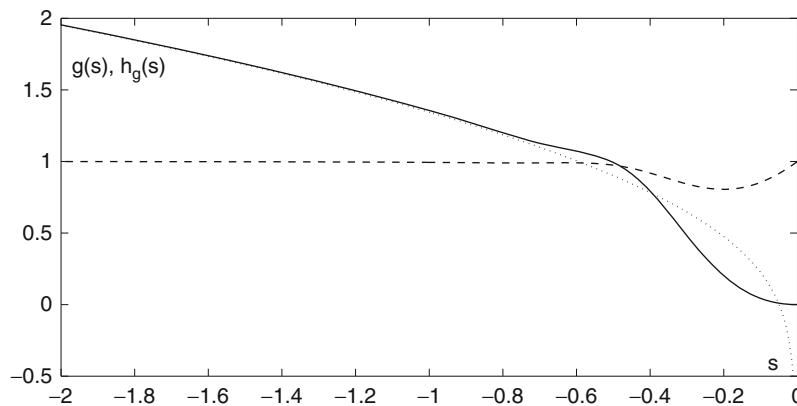
Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Table 1 Dependence of coefficient A and gap $h_q(\Delta r/2)$ at the inlet point on the inlet coordinate α_1

α_1	A	$h_q(\Delta r/2)$
0	0	1
-0.015	0.1	1.07
-0.065	0.2	1.23
-0.154	0.3	1.51
-0.34	0.4	2.15
-1.03	0.5	5.26
-1.98	0.525	11.26
-4.32	0.535	32.96

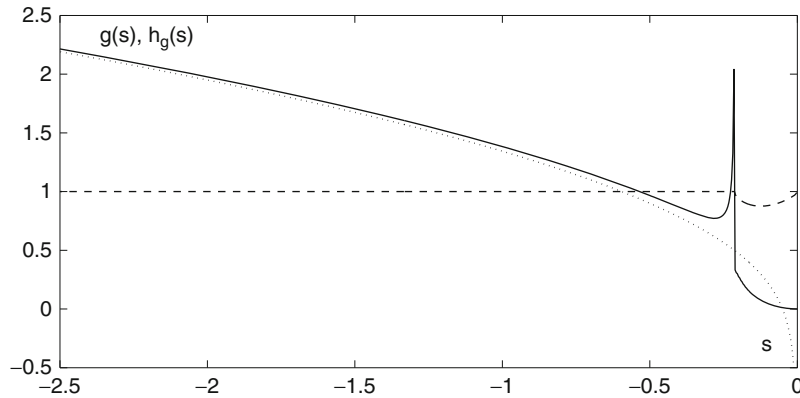
a mild “hump.” However, for larger values of Q_0 a pressure spike is present. It increases in height and becomes very thin as Q_0 increases. This pressure and gap behavior is illustrated in Figs. 6 and 7, the data for which are obtained for $A = 0.525$, $\varepsilon = 0.001$, $Q_0 = 2.5$, and $Q_0 = 10$, respectively. Here, the general pressure behavior is similar to the one of under starved lubrication conditions.

Numerical Solutions of Asymptotic Equations for non-Newtonian Fluids in Pre-critical Lubrication Regimes

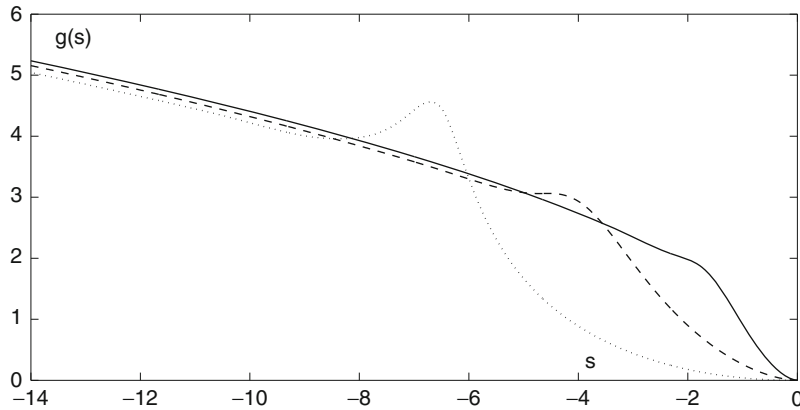
Consider some examples of power law fluid under pure rolling conditions (see Example 2 above) for regimes of starved lubrication for $\varepsilon_q^{1/2} = Q^{-1} \ll V^{1/5}$, $Q \gg 1$ and exponential viscosity $\mu = \exp(Qp)$. In such a case the viscosity in the inlet and exit zones has the form $\mu_q(q) = \exp(Q_0 q)$ and $\mu_g(g) = \exp(Q_0 g)$ for $Q_0 = 1$. The absolute error of calculations is chosen to be $\varepsilon = 0.0001$. All calculations are done for $N = 800$, $\Delta r = 0.03125$ and $N = 2,000$, $\Delta r = 0.0078125$ in the inlet and exit zones, respectively. For example, for $A = 3$ and $n = 0.75$, $n = 1$, and $n = 1.25$ the values of α_1 are equal to -2.3095 , -2.5887 , -2.8163 and the values of β_1 are equal to 1.2908 , 1.5017 , 1.5950 , respectively. For the above values of n , functions $q(r)$ are monotonically increasing with r and approach their asymptotes $q_a(r)$ as $r \rightarrow \infty$. Qualitatively and quantitatively the behavior of pressure distribution $q(r)$ for non-Newtonian fluids ($n \neq 1$) is very similar to the one for Newtonian lubricants ($n = 1$), which is illustrated in Fig. 2. For comparison, three graphs of functions $g(s)$ are given for $A = 3$ in Fig. 8, two of which are determined for non-Newtonian fluid with $n = 0.75$ (solid curve) and $n = 1.25$ (dotted curve), while the third



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 6 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve), and the Hertzian pressure asymptote $g_a(s) = \sqrt{-2s} - \frac{\beta_1}{\sqrt{-2s}}$ (dotted curve) in the exit zone of a fully flooded lubricated contact for $A = 0.525$ and $Q_0 = 2.5$



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 7 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve), and the Hertzian pressure asymptote $g_a(s) = \sqrt{-2s} - \frac{\beta_1}{\sqrt{-2s}}$ (dotted curve) in the exit zone of a fully flooded lubricated contact for $A=0.525$ and $Q_0=10$

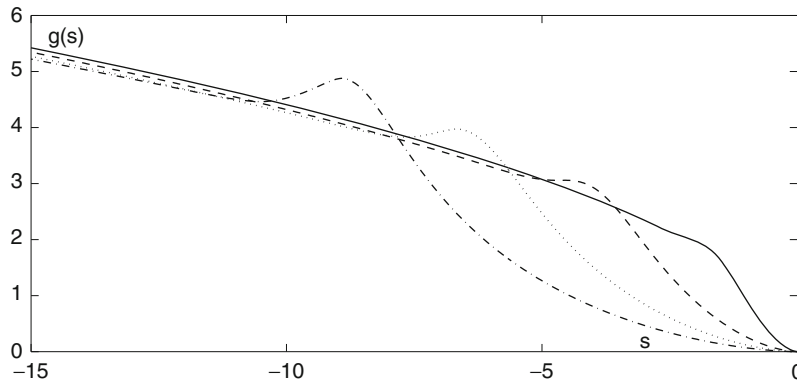


Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 8 Main term of the asymptotic of the pressure distribution $g(s)$ in the exit zone of a starved lubricated contact for $A=3$ and $n=0.75$ (solid curve), $n=1$ (dashed curve), and $n=1.25$ (dotted curve)

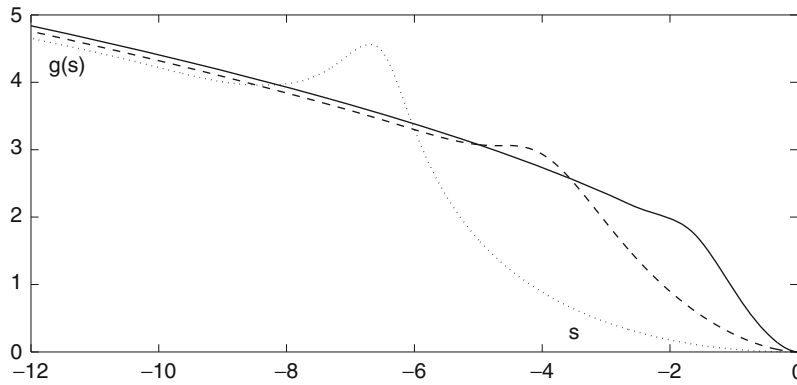
one is determined for the case of Newtonian fluid with $n=1$ (dashed curve). Note that β_1 increases monotonically with n and the pressure spike increases in value and moves toward the contact center as n increases. The behavior of $g(s)$ for large s resembles the behavior of its asymptote $g_a(s)$ and does not exhibit any oscillations or signs of instability.

Figures 9 and 10 demonstrate the behavior of three curves of $g(s)$ for $Q_0=1$, $n=0.75$, and $n=1.25$ (solid, dashed, and dotted curves correspond to $A=1, 2, 3$, respectively). In addition, in Fig. 9 for $n=0.75$ and $A=4$ the pressure distribution $g(s)$ in the exit zone is given by a dash-dotted curve. For $n=1.25$ and $A=4$ the pressure distribution $g(s)$ exhibits some signs of instability and the

problem has to be solved in the regularized form (see ► [Numerical Stability and Precision in Elastohydrodynamic Lubrication \(EHL\)](#) and Kudish and Covitch (2010)). For $n=0.75$ and $A=1, 2, 3$, and 4 it is obtained that $\alpha_1=-0.4619, -1.2848, -2.3095, -3.4761$ and $\beta_1=0.2879, 0.7509, 1.2908, 1.8803$, respectively, while for $n=1.25$ and $A=1, 2, 3$, and 4 the solutions are $\alpha_1=-0.5810, -1.5850, -2.8163, -4.2097$ and $\beta_1=0.3850, 0.9510, 1.5950, 2.6671$, respectively. For $n=0.75$, $A=1, 2$, and $n=1.25$, $A=1$, the pressure distributions $g(s)$ in the exit zone are monotonic while for $n=0.75$, $A=3, 4$ and $n=1.25$, $A=2, 3, 4$, the pressure distributions $g(s)$ possess spikes that shift closer to the center of the lubricated contact (to $s=-\infty$) and increase in value as $|\alpha_1|$



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 9 Main term of the asymptotic of the pressure distribution $g(s)$ in the exit zone of a starved lubricated contact for $n=0.75$ and $A=1$ (solid curve), $A=2$ (dashed curve), $A=3$ (dotted curve), and $A=4$ (dash-dotted curve)

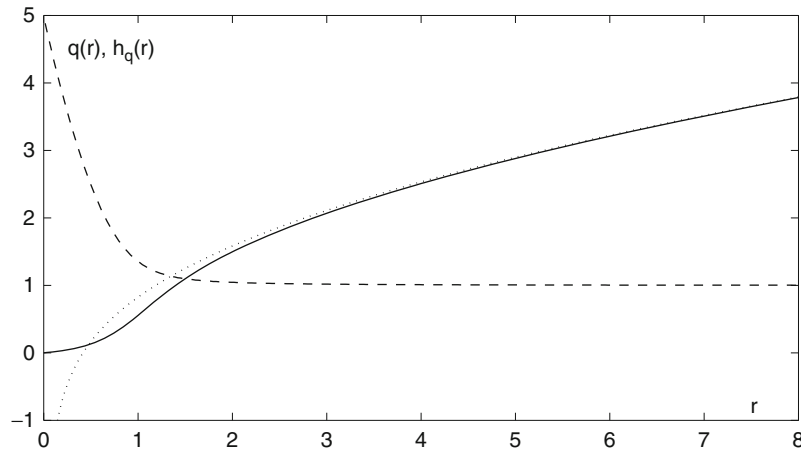


Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 10 Main term of the asymptotic of the pressure distribution $g(s)$ in the exit zone of a starved lubricated contact for $n=1.25$ and $A=1$ (solid curve), $A=2$ (dashed curve), and $A=3$ (dotted curve)

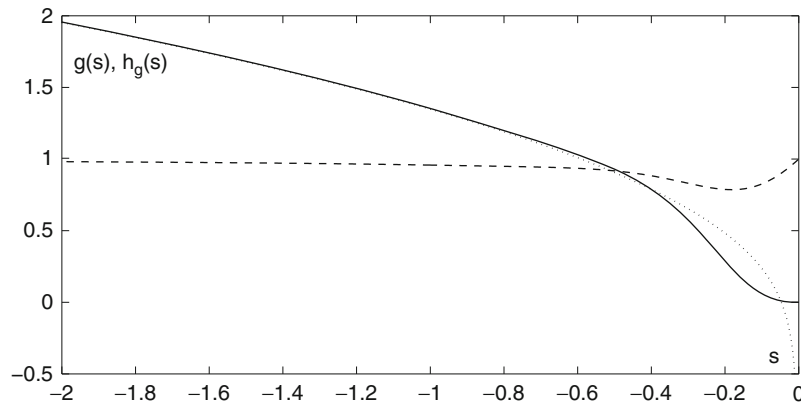
and A increase. As parameter n increases the height of the pressure spike increases and it shifts closer to the contact center. For large s , the behavior of $g(s)$ resembles the behavior of its asymptote $g_a(s)$ and does not exhibit any oscillations or signs of instability.

Consider some results for pre-critical fully flooded lubrication regimes. Qualitatively, the behavior of pressure $q(r)$ and $g(s)$ and gap $h_q(r)$ and $h_g(s)$ is similar to the one for Newtonian fluids described above. For $A=0.4$, $Q_0=1$, and $n=1.25$ ($\alpha_1=-0.8345$, $h_q(\Delta r/2)=4.8837$, $\beta_1=0.0858$, $\min h_g(s)=0.7860$) and $A=0.4$, $Q_0=5$, and $n=1.25$ ($\alpha_1=-0.2395$, $h_q(\Delta r/2)=1.5998$, $\beta_1=0.0610$, \min

$h_g(s)=0.9859$) examples of such solution behavior in the inlet and exit zones are given in Figs. 11, 12, 13 and 14, respectively. In the inlet zone the solutions are obtained for $N=1,200$ and $\Delta r=0.015625$, while in the exit zones for $N=480$ and $\Delta r=0.0078125$. The absolute precision is $\varepsilon=10^{-4}$. To get a better understanding of the solution behavior in the exit zone, Figs. 15 and 16 present graphs of pressure $g(s)$ and gap $h_g(s)$ for four series of input parameters: $A=0.4$, $n=0.75$, $Q_0=5$, and $Q_0=10$ ($\beta_1=0.0496$, $\max g(s)=0.7821$, $\min h_g(s)=0.8967$, and $\beta_1=0.0397$, $\max g(s)=0.8537$, $\min h_g(s)=0.9222$, respectively) and $A=0.4$, $n=1.25$, $Q_0=5$, and $Q_0=10$ ($\beta_1=0.0647$, $\max g(s)=0.8850$,



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 11 Main terms of the asymptotic distributions of pressure $q(r)$ (solid curve), gap $h_q(r)$ (dashed curve), and the Hertzian pressure asymptote $q_a(r) = \sqrt{2r} + \frac{\alpha_1}{\sqrt{2r}}$ (dotted curve) in the inlet zone of a fully flooded lubricated contact for $A=0.4$, $Q_0=1$, and $n=1.25$

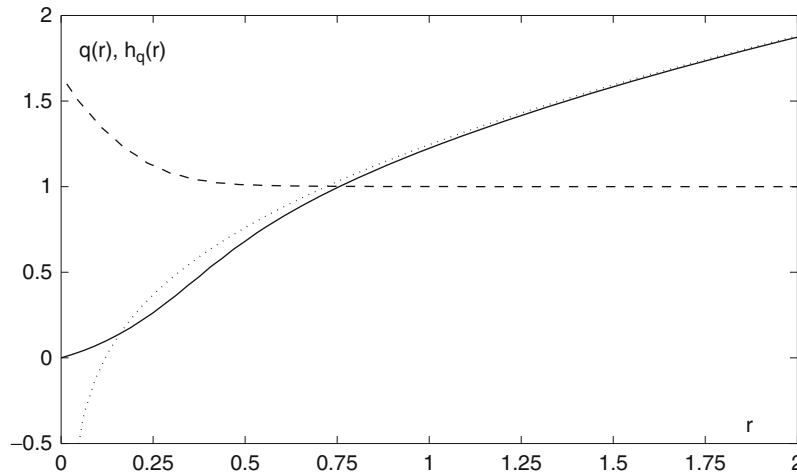


Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 12 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve), and the Hertzian pressure asymptote $g_a(s) = \sqrt{-2s} - \frac{\beta_1}{\sqrt{-2s}}$ (dotted curve) in the exit zone of a fully flooded lubricated contact for $A=0.4$, $Q_0=1$, and $n=1.25$

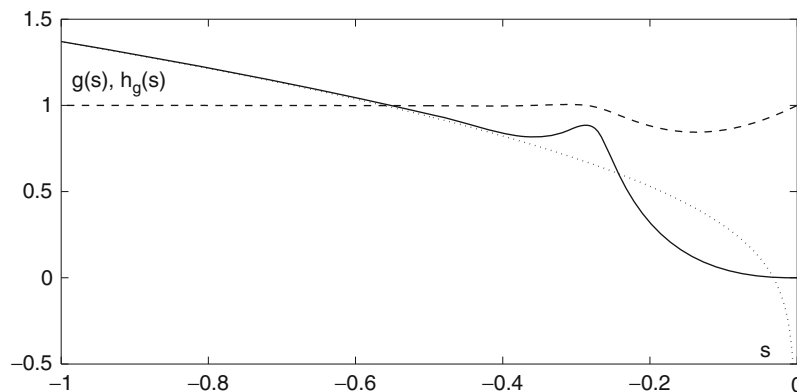
$\min h_g(s)=0.8487$, and $\beta_1=0.08209$, $\max g(s)=2.2897$, $\min h_g(s)=0.877$, respectively). Obviously, pressure $g(s)$ and gap $h_g(s)$ distributions are smoother for $n=0.75$ than for $n=1.25$. The presence of a pin-like pressure spike for $n=1.25$ and $Q_0=10$ is a manifestation of mild instability that can be easily treated by regularization (see ► [Numerical Stability and Precision in Elastohydrodynamic Lubrication \(EHL\)](#) and Kudish and Kovitch (2010)). A similar behavior of pressure $g(s)$ and gap $h_g(s)$ can be observed for $A=0.4$, $n=1.25$, $Q_0=20$, $N=1,200$, $\Delta r=0.00168$

($\beta_1=0.043$, $\max g(s)=2.236$, $\min h_g(s)=0.92$), the graphs of which are depicted in Fig. 17.

Qualitatively, there are significant differences between these two series of solutions. Obviously, for smaller values of Q_0 the sizes of the inlet and exit zones as well as the inlet gap $h_q(\Delta r/2)$ are much wider than for larger values of Q_0 . It can be clearly seen from the comparison of graphs and effective zone sizes in Figs. 11 and 13 in the inlet zones and in Figs. 12 and 14 in the exit zones, respectively. In the exit zone, $\min h_g(s)$ is smaller and located farther from the exit



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 13 Main terms of the asymptotic distributions of pressure $q(r)$ (solid curve), gap $h_q(r)$ (dashed curve), and the Hertzian pressure asymptote $q_a(r) = \sqrt{2}r + \frac{\beta_1}{\sqrt{2r}}$ (dotted curve) in the inlet zone of a fully flooded lubricated contact for $A=0.4$, $Q_0=5$, and $n=1.25$



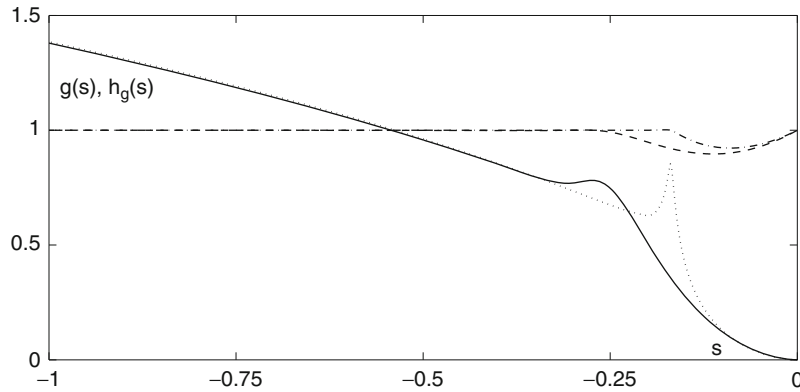
Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 14 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve), and the Hertzian pressure asymptote $g_a(s) = \sqrt{-2}s - \frac{\beta_1}{\sqrt{-2s}}$ (dotted curve) in the exit zone of a fully flooded lubricated contact for $A=0.4$, $Q_0=5$, and $n=1.25$

point for smaller values of Q_0 . Moreover, for smaller values of Q_0 the pressure distribution $g(s)$ in the exit zone is monotonic, while for larger Q_0 it has a spike/local maximum. The value of this pressure spike increases and it shifts toward the center of the lubricated contact as Q_0 increases.

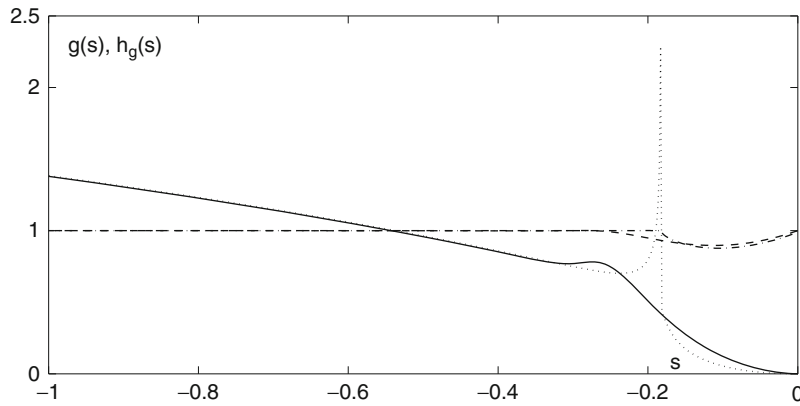
Key Applications

Solutions for starved heavily loaded lubricated contacts are important in high-speed/sliding gears and

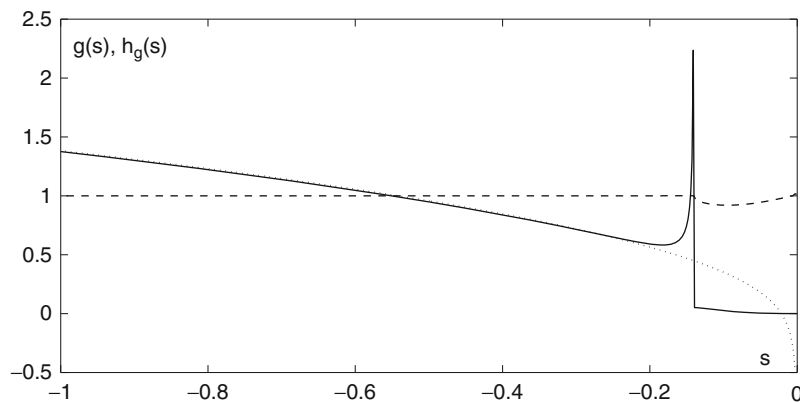
bearings. In most cases, the lubricant behavior is non-Newtonian. The asymptotic approach allows for determining the structure of the EHL problem solution and simple analytical derivation of formulas for the lubrication film thickness and sliding frictional stress. These formulas represent easily obtainable but valuable information about the EHL contact and they can be used in numerical and experimental studies of heavily loaded contacts lubricated by fluids with different non-Newtonian rheologies.



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 15 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve) for $Q_0=5$, and of pressure $g(s)$ (dotted curve), gap $h_g(s)$ (dash-dotted curve) for $Q_0=10$, in the exit zone of a fully flooded lubricated contact, $A=0.4$, $n=0.75$



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 16 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve) for $Q_0=5$, and of pressure $g(s)$ (dotted curve), gap $h_g(s)$ (dash-dotted curve) for $Q_0=10$, in the exit zone of a fully flooded lubricated contact, $A=0.4$, $n=1.25$



Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts, Fig. 17 Main terms of the asymptotic distributions of pressure $g(s)$ (solid curve), gap $h_g(s)$ (dashed curve), and the Hertzian pressure asymptote $g_a(s) = \sqrt{-2s} - \frac{\beta_1}{\sqrt{-2s}}$ (dotted curve) in the exit zone of a fully flooded lubricated contact for $A=0.4$, $Q_0=20$, and $n=1.25$

Cross-References

- [Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts](#)
- [Numerical Stability and Precision in Elastohydrodynamic Lubrication \(EHL\)](#)
- [Stress-Induced Lubricant Degradation and Viscosity Loss](#)
- [Thermoelastohydrodynamically Lubricated Contacts with Non-Newtonian Lubricants: Asymptotic Approach](#)

References

- I.I. Kudish, Asymptotic analysis of a plane non-isothermal elastohydrodynamic problem for a heavily loaded rolling contact. *Proc. Acad. Sci. Armen. SSR Mech.* **31**(6), 16–35 (1978)
- I.I. Kudish, Asymptotic methods of study for plane problems of the elastohydrodynamic lubrication theory in heavy loaded regimes. Part 1. Isothermal problem. *Proc. Acad. Sci. Armen. SSR Mech.* **35**(5), 46–64 (1982)
- I.I. Kudish, Asymptotic method of study for plane problems of the elastohydrodynamic lubrication theory for heavily loaded regimes. Part 2. Non-isothermal problem. *Proc. Acad. Sci. Armen. SSR Mech.* **36**(5), 47–59 (1983)
- I.I. Kudish, Elastohydrodynamic problems for rough bodies with non-Newtonian lubrication. *Dopovidi Akademii Nauk Ukrain's'koi RSR*, 1979, Seriya A, No. 11, pp. 915–920
- I.I. Kudish, M.J. Covitch, *Modeling and Analytical Methods in Tribology* (Chapman & Hall/CRC Press, Boca Raton, 2010)
- I.I. Vorovich, V.M. Aleksandrov, V.A. Babeshko, *Non-classical Mixed Problems of Elasticity* (Nauka Publishing, Moscow, 1974)

Starved Heavily Loaded EHL Contacts

- [Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts](#)

Static Friction of Fibers

- [Fiber Boundary Tribology, Principles](#)

Statistical Contact

- [Stochastic Contact Theories: Other Theories Based on the Greenwood-Williamson Model](#)

Statistical Fracture Mechanics Approach to Contact Fatigue

ILYA I. KUDISH

Department of Mathematics, Kettering University, Flint, MI, USA

Synonyms

[Contact fatigue life](#); [Fatigue crack propagation](#); [Fracture mechanics](#); [Pitting](#)

Definition

Limited contact fatigue (pitting) life of machines and mechanisms is a result of slow material deterioration caused by growth of fatigue cracks. Contact fatigue life is of statistical nature. Scattering of material fatigue lives depends on many parameters, a major one of which is material defectiveness, which manifests itself by the presence of inclusions, voids, and small cracks. In addition, contact fatigue life is affected by material elastic and fatigue properties as well as by applied stresses.

Scientific Fundamentals

An extensive review of more than 60 papers with experimental and theoretical data (Kudish and Burris 2000a) indicates that the basic principles of a sound mathematical model of contact fatigue should be as follows: (1) the existence of the initial statistical defect distribution in materials should be assumed, (2) the existing material defects such as voids and inclusions may be replaced by cracks of equivalent size and orientation, (3) the period of crack initiation may be assumed to be much shorter than the crack propagation period, (4) the changes of the statistical crack distribution due to fatigue crack growth in time should be reflected in the model and should be determined for any time moment, (5) the model should involve the material stress analysis under the action of contact normal and tangential stresses as well as the residual stress, (6) the normal and tangential stress intensity factors at crack tips caused by the stress field should be determined, (7) the subsurface and surface cracks should be taken into account, (8) the fatigue crack growth should be considered according to the Paris equation for fatigue crack growth or a similar equation, (9) the probability of the fatigue damage should be calculated at every material point at any time moment, and (10) the probability of pitting (or the probability of survival of a solid as a whole) should be formulated based on the above-mentioned local probabilities of fatigue damage. Moreover, a review and an

analysis (Kudish and Burris 2000b) of the existing contact fatigue models reveal serious deficiencies of these models. The goal of this entry is to describe and to analyze a fracture mechanics-based model of contact fatigue which takes into account material defectiveness, material elastic and fatigue properties, as well as applied stresses.

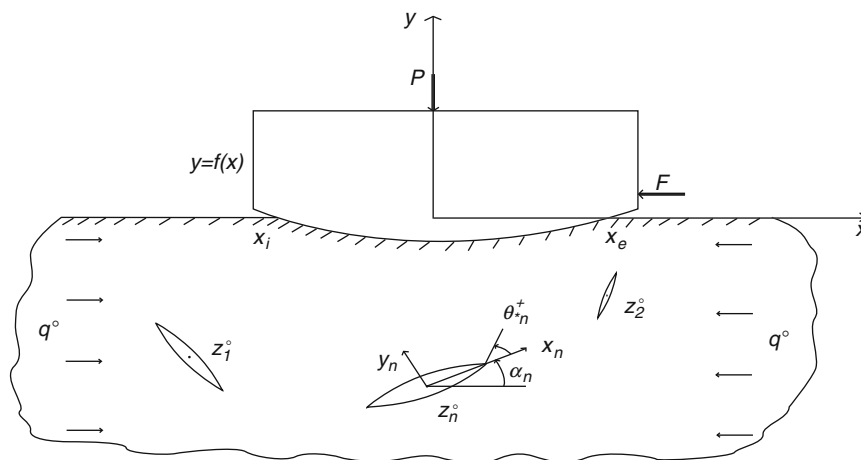
Two-Dimensional Statistical Model of Contact Fatigue

A detailed derivation of a physically sound 2D statistical model of contact fatigue life and its qualitative analysis are given in Kudish (2000). The model is different from the previously published models of contact fatigue (Ioannides and Harris 1985; Tallian 1992a, b, etc.) in several important aspects, such as statistical treatment of material defect distribution which changes in time, stress analysis, crack propagation versus crack initiation, etc. The statistical model of contact fatigue is based on contact and fracture mechanics and a statistical treatment of the initial distribution of material defects. The model takes into account normal, frictional, and residual stresses; initial statistical distribution of defects versus their size, location, and orientation; material fatigue resistance parameters; etc.

Experiments have demonstrated that contact fatigue in steels is due to the presence of defects such as nonmetallic inclusions, carbides, etc. Murakami et al. (1985) have stated that in their tests, all fatigue failures are initiated at subsurface nonmetallic inclusions. A theoretical analysis (Murakami 1987) of stress intensity factors at crack and inclusion tips shows that with high precisions, inclusions can be replaced by cracks of practically the same size. Therefore, the whole contact fatigue model development is based on the analysis of crack behavior. Cracks are

modeled by straight cuts of finite length. Direct calculations (Kudish 2000) show that linear fracture mechanics is applicable during practically the whole life of fatigue cracks except for the very short period of crack explosive growth at the end of their life cycle. It has been reported (Kudish 2002; Kudish and Burris 2004) that under certain operating conditions in the presence of lubricant, the normal stress intensity factor k_1 at the subsurface tips of surface cracks can be several orders of magnitude higher than the one at the tips of similar subsurface cracks. These high values of the normal stress intensity factor k_1 lead to extremely fast propagation of surface cracks in comparison with the subsurface ones, i.e., to very short fatigue life. Under normal operating conditions, fatigue life of bearings and gears can be classified as a high-cycle fatigue phenomenon. Therefore, long fatigue lives observed in tests under normal operating conditions and short fatigue lives resulting from surface-initiated fatigue cracks in the presence of lubricant are irreconcilable (Kudish 2002; Kudish and Burris 2004). It means that in bearings and gears with normally long fatigue life, cracks are initiated at subsurface defects. In practice, material defects are small in comparison to the contact size, they are far from each other and from the material surface (subsurface defects), and they do not interact.

Consider a solid made of an elastic material subjected to cycling loading. Let (x, y, z) be coordinates of a material point in the coordinate system in which the x -axis is directed along the solid surface and is pointed in the direction opposite to material motion, the y -axis is directed inside the material, and the z -axis is perpendicular to the x - and y -axes (see Fig. 1). Suppose $f(0, x, y, z, l_0)$ is the density of the initial distribution of



Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 1 Schematic view of the contact

cracks such that $f(0, x, y, z, l_0)dl_0dxdydz$ is the number of cracks with the half-length between l_0 and $l_0 + dl_0$ in the material volume $dxdydz$ centered at point (x, y, z) with dx , dy , and dz dimensions along the respective axes. The characteristic linear size of volume $dxdydz$ is considered to be much greater than the typical size of the material inhomogeneity and, at the same time, much smaller than the characteristic size of the loaded contact region. Therefore, it is assumed that the defect population in any of the parallelepipeds $dxdydz$ is large enough to ensure an adequate and accurate representation of the phenomenon based on the probabilistic function f . Therefore, the defect distribution $f(0, x, y, z, l_0)$ is assumed to be discrete in the material. The contact fatigue model can be developed for any particular initial distribution $f(0, x, y, z, l_0)$. However, it assumes a simple form if the initial defect distribution $f(0, x, y, z, l_0)$ versus defect initial half-length l_0 is taken as a log-normal distribution with the mean value μ_{ln} and standard deviation σ_{ln} :

$$f(0, x, y, z, l_0) = 0 \text{ if } l_0 \leq 0, \\ f(0, x, y, z, l_0) = \frac{\rho(0, x, y, z)}{\sqrt{2\pi\sigma_{ln}l_0}} \exp\left[-\frac{1}{2}\left(\frac{\ln(l_0) - \mu_{ln}}{\sigma_{ln}}\right)^2\right] \text{ if } l_0 > 0. \quad (1)$$

In (1), $\rho(0, x, y, z)$ is the crack volume density at the initial time moment $N = 0$.

To make further analysis simpler, it is assumed that contact fatigue is initiated only at subsurface defects while there are no defects in a thin surface layer of the material. Under these assumptions at the tips of a small straight subsurface crack with half-length l , the normal (k_1) and shear (k_2) stress intensity factors can be approximately calculated with the help of asymptotic formulas (Kudish 1987) as follows:

$$k_1 = \sqrt{l}[Y^r + q^0 \sin^2 \alpha] \theta[Y^r + q^0 \sin^2 \alpha], \\ k_2 = \sqrt{l}[Y^i - \frac{q^0}{2} \sin 2\alpha], \\ Y = \frac{1}{\pi} \int_{-a_H}^{a_H} [q(t) \overline{D}_0(t) + \tau(t) \overline{G}_0(t)] dt, \tau = -\lambda q, \\ \{Y^r, Y^i\} = \{Re(Y), Im(Y)\}, \\ D_0(t) = \frac{i}{2} \left[-\frac{1}{t-X} + \frac{1}{t-\overline{X}} - \frac{e^{-2ix}(\overline{X}-X)}{(t-\overline{X})^2} \right], \\ G_0(t) = \frac{1}{2} \left[\frac{1}{t-X} + \frac{1-e^{-2ix}}{t-\overline{X}} - \frac{e^{-2ix}(t-X)}{(t-\overline{X})^2} \right], X = x + iy, \quad (2)$$

where $q(x)$ and $\tau(x)$ are contact pressure and frictional stress, q^0 is the residual stress acting along the x -axis, λ is the coefficient of friction, a_H is the Hertzian half-width of the contact, α is the angle between the crack and the positive direction of the x -axis, i is the imaginary unit ($i^2 = -1$), $\theta(\zeta)$ is a step function ($\theta(\zeta) = 0, \zeta \leq 0$ and $\theta(\zeta) = 1, \zeta > 0$), and symbols Re and Im mean the real and imaginary parts of the corresponding quantity. It is important to mention that according to (2), for subsurface cracks the quantities of $k_{10} = k_1 l^{-1/2}$ and $k_{20} = k_2 l^{-1/2}$ are functions of x and y and are independent from l .

In experimental studies of different steels, including bearing ones, it was established by Han and Yang (1987) and Nisitani and Goto (1984) that the crack initiation stage is much shorter than the crack propagation stage. Therefore, soon after cycling loading starts, small fatigue cracks are initiated near inclusions (Kudish and Burris 2000a).

The resultant stress field in a material is formed by the interaction of three types of stresses: stress produced by the contact pressure, frictional stress, and residual stress. Under the action of contact pressure, residual stress, and small frictional stress, the subsurface stress field is dominated by compressive stresses. However, due to the presence of the frictional and tensile residual stresses, the resultant stress may be tensile in the regions behind the contact with respect to the direction of the friction force and in the zones where the residual stress is tensile (Kudish and Burris 2000a, b). Such regions are subsurface and located outside of the contact.

Soon after initiation, fatigue cracks propagate in the direction determined by the local stress field, namely, perpendicular to the local maximum tensile stress (Kudish 2000). Along this direction, the shear stress intensity factor $k_2 = 0$ (Kudish and Burris 2000a), and to find the angle of crack propagation α it is necessary to solve the equation

$$k_2(N, l, \alpha, x, y, z) = 0, \quad (3)$$

where N is the number of loading cycles. The fact that for small subsurface cracks $k_{20} = k_2 l^{-1/2}$ is independent from l together with (3) leads to the conclusion that for cycling loading with constant amplitude, the angle of crack growth α is independent from N . According to (3), at any point (x, y, z) there are two angles α_1 and α_2 along which a crack may propagate. The actual direction of crack propagation α is determined by one of these two angles, α_1 and α_2 , for which the value of the normal stress intensity factor $k_1(N, l, \alpha, x, y, z)$ is greater. If angle α is determined this way, it guarantees that cracks propagate in the direction perpendicular to the maximum tensile stress.

Propagation of fatigue cracks is caused by small tensile stresses resulting from a superposition of pressure and frictional stress as well as from the residual stress. The simplest way to consider propagation of fatigue cracks is to use the Paris equation (Romaniv et al. 1990):

$$\frac{dl}{dN} = g_0 \left(\max_{-\infty < x < \infty} \Delta k_1 \right)^n, l|_{N=0} = l_0, \quad (4)$$

where g_0 and n are the parameters of material fatigue resistance and l_0 is the crack initial half-length.

For constant amplitude of loading taking into account that for subsurface cracks $k_{10} = k_1 l^{-1/2}$ is independent from l , the solution of (4) can be represented in the form (under normal contact conditions $\Delta k_{10} = k_{10}$)

$$l = l_0 \left\{ 1 - N \left(\frac{n}{2} - 1 \right) g_0 \left[\max_{-\infty < x < \infty} k_{10} \right]^n / l_0^{\frac{2-n}{2}} \right\}^{\frac{2}{2-n}}, \quad n > 2. \quad (5)$$

The number of loading cycles needed for a crack to reach its critical length is almost independent from the material fracture toughness K_f . The fact that the small crack propagation phase represents the main period of crack growth allows the use of formulas (2) for stress intensity factors. Formulas (2) were derived for constant residual stresses q^0 . Further on, an additional assumption is used that formulas (2) also can be used in the case of the residual stress q^0 varying slowly with the depth y beneath the surface, i.e., $q^0 = q^0(y)$.

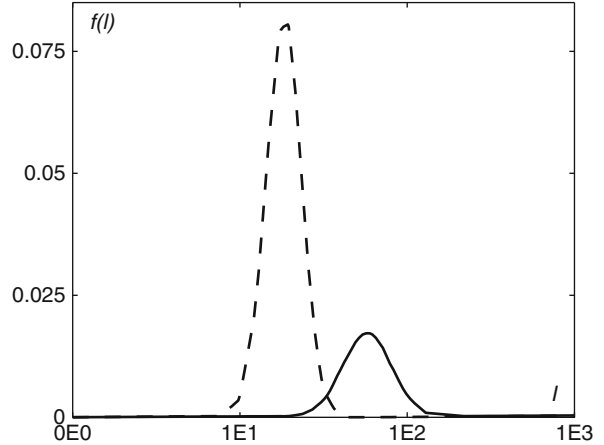
To describe crack statistics after the crack initiation phase is over, it is necessary to make certain assumptions such as the existing cracks do not heal and new cracks are not created. These assumptions hold well due to the fact that the initial material defects are scarce and small (Kudish and Burris 2000b). For constant amplitude of loading, cracks remain small and it can be assumed that over almost all life span of fatigue cracks their orientation does not change. Let $f(N, x, y, z, l)$ be the density of crack distribution as a function of crack half-length l after N loading cycles in a small parallelepiped $dx dy dz$ with the center at (x, y, z) . Then the assumption that the number of cracks in any material volume is constant in time leads to the kinetic equation (Kudish 2000)

$$f(N, x, y, z, l) dl = f(0, x, y, z, l_0) dl_0, \quad (6)$$

which being solved for $f(N, x, y, z, l)$ gives

$$f(N, x, y, z, l) dl = f(0, x, y, z, l_0) \frac{dl_0}{dl}, \quad (7)$$

where l_0 and dl_0/dl as functions of N and l can be obtained from the solution of (5). Equations (5) and (7) give the expression for f after N loading cycles (see Fig. 2):



Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 2 Schematic view of the evolution of the crack distribution $f(N_1, x, y, z, l)$ versus l with time N for $N = N_1$ (dashed curve) and $N = N_2$ (solid curve), $N_2 > N_1$

$$f(N, x, y, z, l) = f(0, x, y, z, l_0(N, l, y, z)) \left\{ 1 + N \left(\frac{n}{2} - 1 \right) g_0 \left[\max_{-\infty < x < \infty} k_{10} \right]^n l^{\frac{n-2}{2}} \right\}^{\frac{n}{n-2}}, \quad (8)$$

$$l_0 = \left\{ l^{\frac{2-n}{2}} + N \left(\frac{n}{2} - 1 \right) g_0 \left[\max_{-\infty < x < \infty} k_{10} \right]^n \right\}^{\frac{2}{2-n}}.$$

It follows from (6) that $\partial \rho(N, x, y, z) / \partial N = 0$.

The material local survival probability $p(N, x, y, z)$ is a certain monotonic measure of the portion of cracks with half-length l below the critical half-length l_c . By choosing the simplest function of such sort that provides for the basic probability properties and using some substitution (Kudish 2000), the expression for $p(N, x, y, z)$ takes the form

$$p(N, x, y, z) = \frac{1}{\rho} \int_0^{l_c} f(0, x, y, z, l_0) dl_0 \text{ iff } (0, x, y, z, l_0) \neq 0, \\ p(N, x, y, z) = 1 \text{ otherwise,} \quad (9)$$

$$l_{0c} = \left\{ l_c^{\frac{2-n}{2}} + N \left(\frac{n}{2} - 1 \right) g_0 \left[\max_{-\infty < x < \infty} k_{10} \right]^n \right\}^{\frac{2}{2-n}},$$

where l_{0c} is the crack initial half-length, which after N loading cycles reaches the critical size of $l_c = (K_f/k_{10})^2$, and ρ is the initial volume density of cracks. For $n > 2$, the value of l_{0c} is a decreasing function of N . Moreover, $l_{0c}(N, y, z)$ is minimal where $k_{10}(x, y, z)$ is maximal, which, in turn, happens where the tensile stress reaches

its maximum. Therefore, to every material, point (x, y, z) is assigned a certain survival probability. Equation (1) determines the material local survival probability $p(N, x, y, z)$ after N loading cycles as a monotonically decreasing function of N .

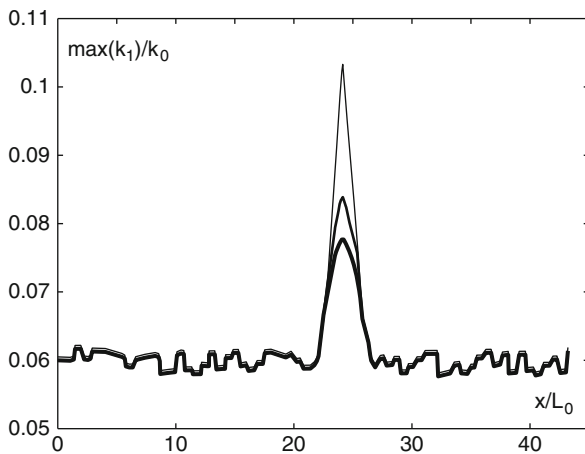
The survival probability $P(N)$ of the material as a whole is determined by local failures. It is assumed that the initial crack distribution in the material is discrete. Then the material survival probability $P(N)$ is equal to

$$P(N) = \prod_{j=1}^{N_c} p(N, x_j, y_j, z_j), \quad (10)$$

where N_c is the total number of cracks in the material stressed volume V . Probability $P(N)$ from (10) satisfies the inequalities

$$[p_m(N)]^{N_c} \leq P(N) \leq p_m(N), \quad p_m(N) = \min_V p(N, x, y, z). \quad (11)$$

Under reasonable assumptions it can be shown (Kudish 2000) that in most cases at the relatively early stages of the fatigue process, $P(N) = p_m(N)$. This is due to the fact that smaller cracks grow slower than the larger ones, which is illustrated in Fig. 3. In Fig. 3 the values of the normal stress intensity factor k_1 are obtained for fatigue cracks randomly distributed over the material volume and are shown after different numbers of loading cycles (k_0 and L_0 are parameters used for scaling). These graphs clearly show that the crack with the initially larger value of the normal stress intensity factor k_1 propagates



Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 3 Illustration of the growth of the initially randomly distributed normal stress intensity factor k_1 with time N

much faster than all other cracks, i.e., the value of its k_1 increases much faster than the values of k_1 for all other cracks which are almost dormant. As a result of that, the crack with the initially larger value of k_1 reaches its critical size way ahead of other cracks. This event determines the time and the place where fatigue occurs initially. Therefore, in spite of formula (10), for high values of n the material survival probability $P(N)$ is a local fatigue characteristic, and it is determined by the material defect with the initially highest value of the stress intensity factor k_1 .

Using (1) and (11), the material survival probability $P(N)$ as a whole is determined by the minimum survival probability $p_m(N)$ (Kudish 2000), i.e.,

$$P(N) = p_m(N) = \frac{1}{2} \left\{ 1 + \operatorname{erf} \left[\frac{\ln \min_V l_{0c}(N, y, z) - \mu_{ln}}{\sqrt{2}\sigma_{ln}} \right] \right\}, \quad (12)$$

where $\operatorname{erf}(x)$ is the error integral. According to (9), for $n > 2$ the value of $\min_V l_{0c}(N, y, z)$ is reached at the point (s) where $\max_V k_{10}$ occurs.

To determine fatigue life N of a contact for the given survival probability $P(N) = P_*$, it is necessary to solve the equation

$$p_m(N) = P_*. \quad (13)$$

Finally, the contact fatigue model is reduced to (2), (3), (12), and (13) for fatigue life N of the material as a whole. Some properties of the model can be found in Kudish (2000).

The relationships between the mean μ_{ln} and standard deviation σ_{ln} of the initial log-normal crack distribution and the regular initial mean μ and standard deviation σ are as follows:

$$\mu_{ln} = \ln \frac{\mu^2}{\sqrt{\mu^2 + \sigma^2}}, \quad \sigma_{ln} = \sqrt{\ln[1 + (\frac{\sigma}{\mu})^2]}. \quad (14)$$

Suppose the material failure occurs at point (x, y, z) with the failure probability $1 - P(N)$. That point determines where in (12) the minimum is reached. At this point in (12), the operation of minimum over the material volume V can be dropped. Solving (12) and (13) leads to the formula for fatigue life:

$$N = \left\{ \left(\frac{n}{2} - 1 \right) g_0 \left[\max_{-\infty < x < \infty} k_{10} \right]^n \right\}^{-1} \left\{ \exp \left[\left(1 - \frac{n}{2} \right) \left(\mu_{ln} + \sqrt{2}\sigma_{ln} \operatorname{erf}^{-1}(2P_* - 1) \right) \right] - \frac{2-n}{l_c^2} \right\}, \quad (15)$$

where $\operatorname{erf}^{-1}(x)$ is the inverse function to the error integral $\operatorname{erf}(x)$. It is possible to simplify this equation for the case

of a material initially free of damage, i.e., when $P(0) = 1$. Discounting the very tail of the initial crack distribution leads to $\max_V l_0 \leq l_c$. Therefore, the second term in l_{0c} dominates the first one. It means that the dependence of l_{0c} on l_c and, consequently, on the material fracture toughness K_f can be neglected. Therefore, (15) can be approximated by

$$N = \left\{ \left(\frac{n}{2} - 1 \right) g_0 \left[\max_{-\infty < x < \infty} k_{10} \right]^n \right\}^{-1} \left\{ \exp \left[\left(1 - \frac{n}{2} \right) (\mu_{ln} + \sqrt{2} \sigma_{ln} \operatorname{erf}^{-1}(2P_* - 1)) \right] \right\}. \quad (16)$$

It follows from (2) that k_{10} is proportional to the maximum Hertzian pressure p_H and, also, depends on the friction coefficient λ and the ratio of the residual stress q^0 and maximum pressure p_H . Making use of (6) and (8) and assuming that $\sigma \ll \mu$, a simple analytical formula for contact fatigue life can be obtained (Kudish 2000):

$$N = \frac{C_0}{(n-2)g_0 p_H^n \mu^{\frac{n}{2-1}}} \exp \left[\left(1 - \frac{n}{2} \right) \frac{\sqrt{2}\sigma}{\mu} \operatorname{erf}^{-1}(2P_* - 1) \right], \quad (17)$$

where constant C_0 depends on the residual stress and the coefficient of friction.

Actually, fatigue life formula (17) can be represented in the form of the Lundberg-Palmgren formula (Lundberg and Palmgren 1947):

$$N = \frac{C_*}{p_H^n}, \quad (18)$$

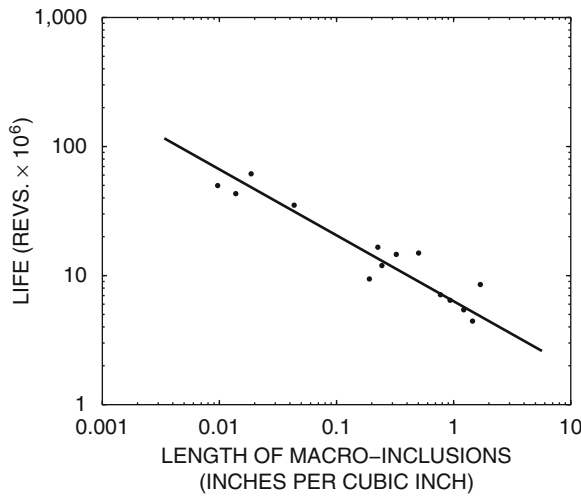
where parameter n can be identified with constant c/e in the Lundberg-Palmgren formula (Lundberg and Palmgren 1947). Formula (18) shows the usual dependence of fatigue life N on the maximum Hertzian stress p_H . The major difference between the Lundberg-Palmgren formula and formula (10) is the fact that in (10), constant C_* depends on material defect parameters μ, σ , coefficient of friction λ , residual stresses q^0 , and probability of survival P_* in a certain way while in the Lundberg-Palmgren formula constant C (the analog of constant C_*) depends only on the depth z_0 of the maximum orthogonal stress, stressed volume V , and probability of survival P_* .

Formula (17) demonstrates the intuitively obvious fact that contact fatigue life N is inversely proportional to the material crack propagation resistance parameter g_0 . Equation (17) exhibits a usual for roller and ball bearings dependence of fatigue life N on the maximum Hertzian pressure p_H . Well-established experimental data for bearings put n in the range between 20/3 and 9. Keeping in mind that usually $\sigma \ll \mu$, for these values of n , contact

fatigue life N is practically inversely proportional to $\mu^{\frac{n}{2-1}}$. Therefore, fatigue life N is a decreasing function of the initial mean inclusion size μ . This conclusion is valid for any value of the material survival probability P_* and is supported by the experimental data discussed in Kudish and Burris (2000b).

It follows from formulas (2) that the stress intensity factor k_1 decreases as the magnitude of the compressive residual stress q^0 increases and/or the magnitude of the friction coefficient λ decreases. Therefore, it follows from formulas (16) and (17) that C_0 is a monotonically decreasing function of the residual stress q^0 and friction coefficient λ . Numerical simulations of fatigue life show (Kudish 2000) that the value of C_0 is very sensitive to the details of the residual stress distribution q^0 versus depth. As the residual stress distribution q^0 and friction coefficient λ are practically solely dependent on the manufacturing operations (such as heat treatment and surface finishing operations), the value of constant C_0 represents the measure of the manufacturing process quality and stability.

The model is validated by the fact that according to the numerical results, $\ln(N)$ is practically a linear function of μ and σ . This behavior of fatigue life N versus μ is similar to the relationship obtained experimentally by Stover et al. (1987). The other way of model validation is the comparison of calculated fatigue lives for tapered roller bearings with the experimental data obtained by the Timken Company for such bearings and presented in Fig. 4 (Stover et al. 1987). The main simplifying assumption made is that bearing fatigue life can be closely approximated by taking into account only the most loaded contact. The following parameters have been used for calculations: $p_H = 2.12$ GPa, $a_H = 0.265$ mm, $\lambda = 0.002$, $g_0 = 6.009 \text{ MPa}^{-n} \cdot \text{m}^{1-n/2} \cdot \text{cycle}^{-1}$, $n = 6.67$, the residual stress varied from $q^0 = -237.9$ MPa on the surface to $q^0 = 0.035$ MPa at the depth of $400 \mu\text{m}$ below the surface, and fracture toughness K_f varied between 15 and 95 $\text{MPa} \cdot \text{m}^{1/2}$. The crack (inclusion) initial mean half-length μ varied between 49.41 and 244.25 μm ($\mu_{ln} = 3.888 - 5.498 + \ln(\mu\text{m})$) and the crack initial standard deviation varied between $\sigma = 7.61$ and 37.61 μm ($\sigma_{ln} = 0.1531$). The results for fatigue life $N_{15.9}$ (for $P(N_{15.9}) = P_* = 0.159$) calculations are given in Table 1 and practically coincide with the experimental data obtained by the Timken Company and presented in Fig. 4. However, it has to be understood that there are certain differences in the numerically obtained data and the data presented in Fig. 4. In Fig. 4, fatigue life is given as a function of the cumulative inclusion length per cubic inch of steel, and here, fatigue life is calculated as



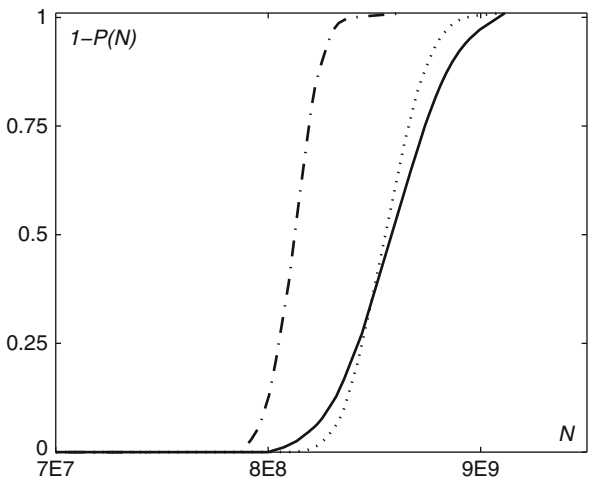
Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 4 Bearing life–inclusion length correlation (Fig. 19 from Stover et al. (1987)) (Reprinted with permission from the Timken Company)

Statistical Fracture Mechanics Approach to Contact Fatigue, Table 1 Relationship between the tapered bearing fatigue life $N_{15,9}$ and the initial inclusion size mean μ and standard deviation σ

$\mu[\mu\text{m}]$	$\sigma[\mu\text{m}]$	$N_{15,9}[\text{cycles}]$
49.41	7.61	2.5×10^8
73.13	11.26	1.0×10^8
98.42	15.16	5.0×10^7
147.11	22.66	2.0×10^7
244.25	37.62	6.0×10^6

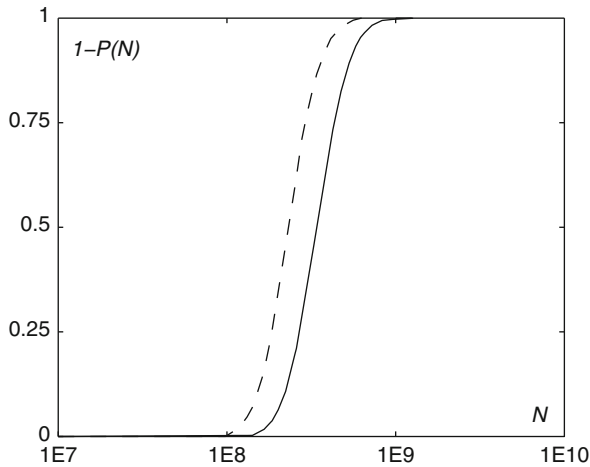
a function of the mean inclusion half-length μ . The numerical data for fatigue life can be brought in perfect compliance with the experimental data by choosing the right measure of steel cleanliness and the proper values for the parameters of the material fatigue resistance g_0 and n .

Consider several examples. Let the maximum Hertzian pressure $p_H = 2\text{GPa}$, contact region half-width in the direction of motion $a_H = 0.249\text{mm}$, friction coefficient $\lambda = 0.002$, residual stress varying from $q^0 = -237.9\text{MPa}$ on the surface to $q^0 = 0.035\text{MPa}$ at the depth of $400\mu\text{m}$ below it, fracture toughness K_f varying between 15 and $95\text{MPa} \cdot \text{m}^{1/2}$, $g_0 = 8.863\text{MPa}^{-n} \cdot \text{m}^{1-n/2} \cdot \text{cycle}^{-1}$, $n = 6.67$, mean of crack initial half-lengths $\mu = 49.41\mu\text{m}$ ($\mu_{ln} = 3.888 + \ln(\mu\text{m})$), and crack initial standard

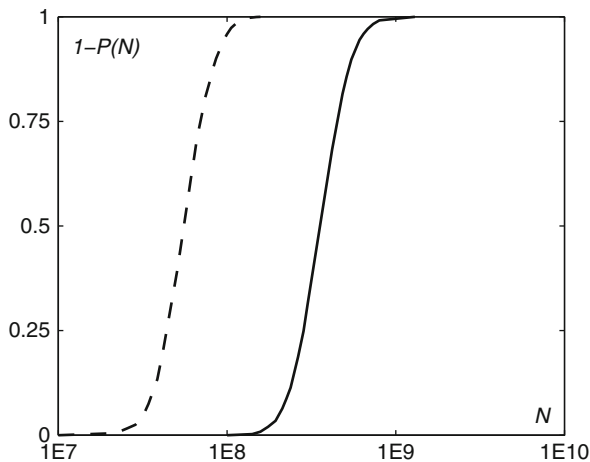


Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 5 Graphs of pitting probability $1 - P(N)$ calculated for the basic set of parameters (solid curve) with $\mu = 49.41\mu\text{m}$, $\sigma = 7.61\mu\text{m}$ ($\mu_{ln} = 3.888 + \ln(\mu\text{m})$, $\sigma_{ln} = 0.1531$), for the same set of parameters and the increased initial value of crack mean half-lengths (dash-dotted curve) $\mu = 74.12\mu\text{m}$ ($\mu_{ln} = 4.300 + \ln(\mu\text{m})$, $\sigma_{ln} = 0.1024$), and for the same set of parameters and the increased initial value of crack standard deviation (dotted curve) $\sigma = 11.423\mu\text{m}$ ($\mu_{ln} = 3.874 + \ln(\mu\text{m})$, $\sigma_{ln} = 0.2282$)

deviation $\sigma = 7.61\mu\text{m}$ ($\sigma_{ln} = 0.1531$) be the basic set of model parameters typical for bearing testing. Numerical results show that fatigue life is practically independent from the material fracture toughness K_f . To illustrate the dependence of contact fatigue life on some of the model parameters, just one parameter from the basic set of parameters will be varied at a time and graphs of the pitting probability $1 - P(N)$ for these sets of parameters (basic and modified) will be compared. Figure 5 shows that as the initial values of the mean μ of crack half-lengths and crack standard deviation σ increase, contact fatigue life N decreases. The effect of increased “friction coefficient” on contact fatigue is shown in Fig. 6. Contact fatigue life decreases as the magnitude of the tensile residual stress and/or friction coefficient increase (see Fig. 7). The results show that fatigue life does not change when the magnitude of the compressive residual stress is increased/decreased by 20% of its base value while the tensile portion of the residual stress distribution remains the same. This conclusion is in agreement with the fact that tensile stresses control fatigue. However, when the compressive residual stress becomes small enough, the acting frictional stress may supersede it and create new regions with tensile



Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 6 Graphs of pitting probability $1 - P(N)$ calculated for the basic set of parameters including $\lambda = 0.002$ (solid curve) and for the same set of parameters and the increased friction coefficient (dashed curve) $\lambda = 0.004$



Statistical Fracture Mechanics Approach to Contact Fatigue, Fig. 7 Graphs of pitting probability $1 - P(N)$ calculated for the basic set of parameters (solid curve) and for the same set of parameters and changed profile of residual stress q^0 (dashed curve) in such a way that at points where q^0 is compressive, its magnitude is unchanged, and at points where q^0 is tensile, its magnitude is doubled

stresses which potentially may cause acceleration of fatigue failure.

This methodology is flexible enough to allow for different modifications. It has been applied to the

engineering analysis of wear and contact fatigue in cases of lubricant contaminated by rigid abrasive particles and contact surfaces charged with abrasive particles (Kudish 1991) as well as to the calculation of bearing wear and contact fatigue life (Kudish 1990).

Being applied to bearings and/or gears, the described statistical contact fatigue model can be used as a research and engineering tool in modeling pitting. In the latter case, some of the model parameters may be assigned certain fixed values based on the scrupulous analysis of steel quality and the quality and stability of manufacturing processes. The model takes into account the most important parameters of the contact fatigue phenomenon (such as normal and frictional contact and residual stresses, initial statistical defect distribution, orientation of fatigue crack propagation, material fatigue resistance, etc.). The model allows for examination of the effect of variables such as steel cleanliness, applied stresses, residual stress, etc., on contact fatigue life as single or composite entities.

This model of contact fatigue has been extended on a three-dimensional case (Kudish 2007), which is also applicable to structural fatigue. In a 3D case, the driving force of crack propagation is the maximum principal tensile stress, while the crack orientation is determined by the plane at which both principal shear stresses are equal to zero. An extension of the model on cases of stochastic loading, variable loading amplitude, and periodic loading regimes can be found in Kudish and Covitch (2010).

Key Applications

Prediction of fatigue life of various mechanisms is extremely important for proper design, exploitation, and maintenance of these mechanisms. In particular, it is of paramount importance for gears and bearings. The model provides a research and practical tool for modeling and understanding of the effect of various parameters (such as initial material defectiveness, material elastic and fatigue properties, residual stress, and applied operating stresses) on material contact fatigue life.

Cross-References

► [Stress-Induced Lubricant Degradation and Viscosity Loss](#)

References

- W.C. Han, C.C. Yang, On the influence of strain-path in multiaxial fatigue failure. *ASME J. Eng. Mater. Technol.* **109**(2), 107–113 (1987)
- E. Ioannides, T.A. Harris, A new fatigue life model for rolling bearings. *ASME J. Tribol.* **107**, 367–378 (1985)

- I.I. Kudish, Contact problem of the theory of elasticity for pre-stressed bodies with cracks. *J. Appl. Mech. Tech. Phys.* **28**, 144–152 (1987)
- I.I. Kudish, Statistical calculation of wear and contact fatigue of rolling bearings. *J. Frict. Wear* **11**(5), 933–944 (1990)
- I.I. Kudish, Wear and fatigue pitting taking into account contaminated lubricants and abrasive particles indented into working surfaces. *J. Frict. Wear* **12**(3), 713–725 (1991)
- I.I. Kudish, A new statistical model of contact fatigue. *STLE Tribol. Trans.* **43**, 711–721 (2000)
- I.I. Kudish, Lubricant-crack interaction, origin of pitting, and fatigue of drivers and followers. *STLE Tribol. Trans.* **45**, 583–594 (2002)
- I.I. Kudish, Fatigue modeling for elastic materials with statistically distributed defects. *ASME J. Appl. Mech.* **74**, 1125–1133 (2007)
- I.I. Kudish, K.W. Burris, Modern state of experimentation and modeling in contact fatigue phenomenon. Part I. Contact fatigue versus normal and tangential contact and residual stresses. Nonmetallic inclusions and lubricant contamination. Crack initiation and crack propagation. Surface and subsurface cracks. *STLE Tribol. Trans.* **43**(2), 187–196 (2000a)
- I.I. Kudish, K.W. Burris, Modern state of experimentation and modeling in contact fatigue phenomenon. Part II. Analysis of the existing statistical mathematical models of bearing and gear fatigue life. New statistical model of contact fatigue. *STLE Tribol. Trans.* **43**(2), 293–301 (2000b)
- I.I. Kudish, K.W. Burris, Modeling of surface and subsurface crack behavior under contact load in the presence of lubricant. *Int. J. Fract.* **125**, 125–147 (2004)
- I.I. Kudish, M.J. Covitch, *Modeling and analytical methods in tribology* (Chapman & Hall/CRC, Boca Raton, 2010)
- G. Lundberg, A. Palmgren, Dynamic capacity of rolling bearings. *Acta Polytechnica (Mech. Eng. Ser. 1)*, R. Swed. Acad. Eng. Sci. **7**, 5–32 (1947)
- Y. Murakami (ed.), *Stress Intensity Factors Handbook*, vol. 1 (Pergamon, Oxford, 1987)
- Y. Murakami, M. Kaneta, H. Yatsuzuka, Analysis of surface crack propagation in lubricated rolling contact. *ASLE Trans.* **28**, 60–68 (1985)
- H. Nisitani, M. Goto, Effect of stress ratio on the propagation of small crack of plain specimens under high and low stress amplitudes. *Trans. Jpn. Soc. Mech. Eng., Ser. A* **50**(453), 1090–1096 (1984)
- O.N. Romaniv, S.Y. Yarema, G.N. Nikiforchin, N.A. Makhutov, M.M. Stadnik, *Fracture mechanics and strength of materials*. Fatigue and cyclic crack resistance of construction materials, vol. 4 (Naukova Dumka, Kiev, 1990)
- J.D. Stover, R.V. Kolarik II, The evaluation of improvements in bearing steel quality using an ultrasonic macro-inclusion detection method. The Timken Company Technical Note (January 1987), pp. 1–12
- T.E. Tallian, Simplified contact fatigue life prediction model – Part I: review of published models. *ASME J. Tribol.* **114**(1), 207–213 (1992a)
- T.E. Tallian, Simplified contact fatigue life prediction model – Part II: new model. *ASME J. Tribol.* **114**(1), 214–222 (1992b)

Statistical Self-Similarity of Discrete Contact

► [Fractal Contact Mechanics](#)

Steam Turbine Lubricants

► [Steam Turbine Oils](#)

Steam Turbine Oils

JAMES B. HANNON

Division of Exxon Mobil Corporation, ExxonMobil Lubricants & Petroleum Specialties Company, Allentown, NJ, USA

Synonyms

R & O oil; Steam turbine lubricants; Turbine oil

Definition

Steam turbine lube oil enables the rotation of large turbine rotors, supported by journal bearings and centered by thrust bearings, in the generation of electricity. Steam turbine oil facilitates reliable rotation of the large turbine rotors by providing two major functions, lubrication and heat removal. The turbine oil must provide suitable lubrication as the turbine rotor speed increases from stationary to low speed where boundary lubrication is the primary lubrication mode, and finally to full speed, where hydrodynamic lubrication is in effect. Maintaining clean, cool, and dry steam turbine oil supports long and reliable turbine operation and turbine oil life.

Scientific Fundamentals

Proper lubrication of steam turbines is critical to avoid costly equipment downtime and repair. The key elements of steam turbine lubrication are:

- Turbine oil selection
- Turbine oil system components and operation
- Turbine oil maintenance

Understanding these three key elements can support years of trouble-free operation, with some steam turbine oils lasting for more than 30 years with nominal make-up rates.

Turbine Oil Selection

High-quality steam turbine oils should provide:

Lubricant characteristic	Operational benefit
Rust and corrosion protection	Protects surface metals
Oxidation stability	Contributes to long oil life
Viscosity stability	Prevents wear, maintains rotor alignment
Demulsibility	Sheds water, minimizes corrosion and bearing wear
Deposit control, to minimize sludge and varnish	Minimizes valve sticking, bearing failure, and filter fouling
Foam control	Minimizes reservoir overflow or level reading issues
Filterability, with retention of additive chemistry	Promotes cleaner systems and maintains oil additive performance
Air release in low oil residence systems	Minimizes micro-dieseling (adiabatic compression) and sluggish valve control
Wear protection, if required in gear-driven generators	Minimizes gear tooth wear

Of these key steam turbine oil characteristics, the most important performance requirements are the ability to (1) protect against rust and corrosion, (2) provide oxidation stability, (3) maintain viscosity, and (4) shed water. Water contamination prevention is the leading reason for steam turbine oil performance properties in these systems.

Steam turbine oils should be formulated for excellent water separation performance as water will promote oil degradation, which ultimately can damage turbine rotating and stationary internal lube circuit components. In steam turbines, water contamination is almost inevitable, requiring maintenance practices, equipment, and the turbine oil to be carefully monitored. Water in oil can reduce bearing life, increase rust and corrosion, lead to sludge formation, and accelerate lubricant oxidation. In some cases, water contamination in warm stagnant areas has led to the formation of microbial growth, which can block filters and control oil systems, degrade oil, and produce corrosive byproducts.

Formulating a steam turbine oil to reliably deliver these critical performance parameters requires carefully selected lubricant base stocks and balanced additive technology. In general, steam turbine oils are approximately

99% base oil and 1% additive. Base stocks can be selected from API base stock groups I, II, III, and IV.

API base stock classification				
Group	Physical specifications			Manufacturing process
	VI	Sulfur (%wt.)	Saturates (%wt.)	
I	90–120	>0.03	<90	Conventional (solvent refined)
II	90–120	<0.03	>90	Hydroprocessed
III	>120	<0.03	>90	Severely hydroprocessed
IV	>130	0.00	>90	Chemical synthesis
V	Vary	Vary	Vary	All other types

Each base stock has a unique set of advantages and disadvantages. In general, higher group base stocks will have better oxidation stability. In contrast, Group I base stocks tend to have better solvency than Groups II and III. Base stock price and availability are also key considerations in product formulation. Base oil oxidation stability and solvency may also be addressed through the selection of additives. Mixture of Group I, II, and III base stocks should not present operational issues as long as the additive technologies are understood by the oil supplier.

Group V base stocks have unique properties that do not follow any specific rule of thumb. An example of a Group V base stock is phosphate ester, which is often used as control oil for its inherent fire-resistant properties.

Steam turbine oil additives are designed to improve the following characteristics:

- Antioxidant – often a mix of additive packages of hindered phenols and aromatic amines are used to support long-term performance and reduced deposit formation.
- Rust and corrosion inhibitors – ashless acids or esters are examples of surface active additives that prevent water access to metal surfaces.
- Demulsifiers – mixed polyol esters or ethers can enhance water separation by promoting water drop coalescence enabling free water separation by gravity-driven phase separation.
- Defoamants – reduce surface tension so that air bubbles can be released. Silicones or polyacrylates are the two common classifications of antifoam additives.
- Metal Passivator – triazoles offer protection from yellow metal corrosion and can provide some level of oxidation resistance.

- Gear wear protection, for gear-driven generators – often phosphorus based, forming a protective and sacrificial layer on the gear tooth.

Thorough blending of these additives into the base oil is needed to ensure a complete and homogeneous blend. Blending involves precise additive dosage in the proper sequence, at a controlled temperature in blend kettles that promote complete mixing. Care must be taken during the blending process to minimize external contamination.

Steam turbine oils are typically blended in ISO viscosity grades of 32, 46, and 68 with ISO grades 32 and 46 being used extensively in large utility-sizes steam turbines. ASTM D 4304 – *Standard Specification for Mineral Lubricating Oil Used in Steam or Gas Turbines* offers new oil selection guidance based on key physical, chemical, and performance criteria. Many steam and gas turbine Original Equipment Manufacturers (OEM) have developed their own acceptance specifications.

Steam Turbine Oil System Components and Operation

To better understand the performance requirements of a steam turbine oil it will be useful to offer system component and operating details. Steam turbines range in size and application from a single-stage, back-pressure turbine driving a pump or compressor to a multiple-stage compound casing, condensing turbine with steam extractions, that can generate up to 1,500 MW. The following details are in reference to large, utility-sized, high-pressure, superheated steam to condensing, horizontal shaft, electric generating turbines.

Large high steam pressure turbines typically rotate at 1,800 or 3,600 rpm for 60 Hz generators and 1,500 or 3,000 rpm for 50 Hz generators. These turbines are typically comprised of three or four separate casings (or shells), high pressure, medium pressure, and low pressure. The low-pressure turbine typically exhausts to a shell and tube condenser, where a vacuum is created as the steam bubbles collapse to water. Establishing and maintaining a high vacuum, targeting 29.92 in. mercury (1 bar), provides optimum thermodynamic efficiencies.

A turbine rotor that can weigh more than 100 t (91 metric tons) is normally supported by two journal bearings per casing. These journal bearings are typically bimetallic, with softer metal bonded to a hard steel shell.

The softer metal material is often babbit comprised of tin alloyed with copper, iron, antimony, and lead. Babbit thickness can range from 1 mil (25 μm) to 100 mil (25,000 μm). A mil is one-thousandths of one inch (0.001 in.). The journal is normally an integral part of the turbine rotor and made from the same material steel. In a resting position, a 20 in. (50.8 cm) diameter journal bearing may have a clearance of 30 mils (760 μm). In operation, an oil wedge of .78–1.57 mils (20–40 μm) is formed. Babbit creep and bearing failure can occur at temperatures above 270F (132°C) at nominal loads of 200–1,000 psi (13.8–68.9 bar). Bearing-embedded thermocouples and/or lubricant drain temperature measurement and alarming are suggested. Bearing metal temperature is typically 30–80°F (17–45°C) higher than the drain oil temperature. Lubricating oil is supplied through drilled passages at pressures ranging from 10 to 20 psig (0.7–1.38 bar) and temperatures ranging from 110°F to 145°F (43–63°C). A lubricating oil typically gains 20–50°F (10.7–27.7°C) between bearing inlet and bearing outlet. Oil outlet temperatures should typically not exceed 180°F (82°C), as elevated temperatures can promote premature degradation of the oil. Proper oil flow to these bearings provides fluid film for hydrodynamic lubrication and heat removal to protect the bearing from bearing overheating. Most bearing oil outlets are equipped with a direct-read thermometer and a sight flow port for visual inspection.

Most large steam turbines will also require a thrust bearing to prevent rotor blade axial misalignment with the stator blades. Severe axial misalignment could result in catastrophic failure should the rotor blade fracture. Minor axial misalignment can result in efficiency losses and damage to rotor blades and seals. Thrust bearings are most often of the self-equalizing tilt pad design. Lubricant is supplied at 10–20 psig (0.7–1.38 bar) pressure and in operation will provide a hydrodynamic film wedge of protection and remove heat. Thrust bearing and oil drain temperatures should be monitored to confirm proper lubrication.

Bearing supply piping is typically surrounded by the return piping, pipe within pipe annular design. This is done as fire hazard prevention in an effort to contain the pressurized, typically 50–60 psig (3.4–4.1 bar) supply pipe in the event of a rupture from spaying combustible lubricant toward hot steam turbine surfaces. Lube oil returns from the bearings in the external pipe within pipe and gravity flow through a series of return screens into the steam turbine oil reservoir.

The turbine oil lubricant is stored in a reservoir, ranging in size up to 20,000 gal (75,700 l), which is typically located on the ground floor, below the steam turbine deck. The reservoir is sized to provide suitable oil residence time that allows separation of deposits, contaminants, and air. The reservoir is kept at a slightly negative pressure, 1–2 in. H₂O (250–500 Pa), by a vapor extractor to minimize potential build-up of dangerous hydrogen gas. Turbine oil at approximately 60 psig (4.1 bar) is pumped from the reservoir to the turbine bearings and is typically maintained at temperatures between 100°F and 120°F (32°C and 43°C) by shell and tube lube oil coolers. Cooling water for these heat exchangers is often supplied from cooling towers or raw water heat exchangers.

During turbine start up and shut down the turbine is typically put on turning gear to minimize temperature-related distortion or bowing of the turbine shaft. The turning gear is typically an electric or hydraulic motor that is coupled to the shaft. At turbine shaft speeds of less than 100 rpm, a separate positive displacement oil pump, often called a jacking oil pump or lift oil pump, will supply oil to the bearings to lift and protect the rotor until sufficient shaft speed is obtained to develop a hydrodynamic film. At start up, an electric motor-driven, vertical shaft, auxiliary lube oil pump is in operation. When the turbine shaft reaches approximately 90% of rotating speed a turbine shaft-driven pump, located in the turbine front standard, will supply full oil flow to journal and thrust bearings and the jacking oil pump will shut down.

Fluid conditioning is typically done via kidney loop from the main turbine oil reservoir. This kidney loop lubricant conditioning process can involve multiple pieces of equipment that work to remove water and insoluble contaminants. In many large steam turbine facilities a separate, mechanical, gravity, settling tank, with bag filters, screens and weir gates are used to precipitate out free water and insoluble contaminants. Often times these mechanical settling tanks are replaced with newer, more effective, oil conditioning equipment. ASTM D 6439 – *Standard guide for Cleaning, Flushing, and Purification of Steam, Gas and Hydroelectric Turbine Lubrication Systems* offers additional general guidance.

Turbine oil conditioning technologies include centrifugation, coalescence, vacuum dehydration, membrane moisture exchange, electrostatic precipitation, and polishing filtration. These oil conditioning technologies can have unique advantages and disadvantages in their use. In general, the water removal

technologies described below offer the following operational performance:

Water removal technology	Operational performance
Centrifugation	Removes free water, not dissolved or emulsified water
Coalescence	Removes free and emulsified water, not dissolved water
Vacuum dehydration	Removes free, emulsified, and dissolved water
Membrane moisture exchange	Removes free, emulsified, and dissolved water

Water removal is a primary concern in steam turbine applications, so care must be taken to confirm that water intrusion is kept at a minimum and that water removal equipment is functioning properly. The primary source of water in a steam turbine is from gland sealing steam that is supplied to bearing gland seals. Gland sealing steam is required to supplement the sealing effectiveness of labyrinth or carbon seals whereas turbine steam, at a positive pressure, would escape into operating spaces, or at a negative pressure air would erode condenser vacuum. Gland sealing steam is typically saturated steam regulated to pressures from 0.5 to 2 psig (0.034–0.14 bar) in the inner gland pockets. Gland sealing steam travels in both directions along the turbine shaft through the seal area exhausting to the steam/condensate leak-off path or entering the casing. Steam channeled through the leak-off can be aided by an exhaustor (blower) or venturi. Gland seal leak-off is kept at a vacuum, typically 10 in. water (2,500 Pa). A fraction of this steam also condenses and commingles with the turbine lube oil in the journal bearing and is carried back to the turbine oil reservoir through the bearing drain return piping. Compromised bearing gland seals or failed exhaustor systems can be the source of excessive water contamination. Also, high gland sealing steam pressure can result in elevated water contamination. Therefore, turbine oil system design and oil conditioners must allow for rapid water removal. At a minimum the turbine oil reservoir should be fitted with a bottom drain, water trap, or “U” tube with a water leg.

To improve the generator's efficiency, large utility generators are often cooled with hydrogen versus air, which provides more effective winding cooling. Hydrogen gas, under pressure, up to 60 psi (4.1 bar) is cooled by plant cooling water in shell and tube heat exchangers mounted on or near the generator casing.

A hydrogen seal oil system, using steam turbine oil, is used in the generator seal to limit hydrogen leakage into work spaces. Oil supplied from the shaft-driven lube oil pump to the seals may be passed through a vacuum tank, which removes air, hydrogen, and water. Return oil is ported back to the main oil reservoir. When the turbine is not rotating, hydrogen seal oil is supplied by a motor-driven pump. Some seal oil systems use a “loop seal” with a vapor extractor for removal of potentially dangerous hydrogen gasses. Excessive seal oil foaming can occur due to water or airborne particulate contamination.

High-pressure superheated steam turbines typically use a separate, high-pressure, 1,500–2,000 psi (103–138 bar) control oil for steam chest valve modulation. Conventional mineral oils are typically not used in this application due to flammability concerns should this oil spray onto the very high temperature surfaces of the steam chest or inlet steam piping. A fire-resistant phosphate ester is often used as the electro hydraulic control (EHC) fluid. The emergency trip valve, sometimes called the overspeed stop valve, will also typically utilize the same phosphate ester fluid. Some low pressure hydraulic systems will use the same oil as used in the bearings and may actually share the same oil reservoir.

Most large utility-sized generators are direct driven from the turbine shaft. In limited numbers and typically for smaller turbine applications the generator may be gear driven. Gear-driven generators typically require a level of gear tooth wear protection in the lubricant. Gear tooth wear performance of the lubricant is normally demonstrated in FZG rig testing, ASTM D 5182 *Test Method for Evaluating the Scuffing Load Capacity of Oils*. A failure load stage of eight or nine is often considered acceptable for gear-driven generators.

Turbine Oil Maintenance

A well-designed oil analysis program, in conjunction with proper turbine oil conditioning equipment, is recommended to support reliable turbine operation. Selected turbine oil tests can be a valuable indicator of turbine oil condition and the turbine itself.

For the highest data quality, a well-positioned sample valve would be installed at the outlet of each bearing drain. This location would allow for bearing-specific trouble shooting. Original design typically does not allow for this and field modifications are rarely installed. Most often a reservoir dip/siphon sample or a pump discharge sample provides for an acceptable turbine oil sample point location. Reservoir drains should not be utilized as they do

not offer a representative sample. Care should be taken not to contaminate the lube oil sample, which is often done in the process of opening or closing the sample valve while filling the sample bottle. Sample oil connections should be well purged to a point where the sample oil approximates operating temperature.

Sample test slate and frequency will be discussed in parallel, addressing frequent field checks, routine analysis, and advanced suitability for continued use programs.

Frequent Field Checks

With the use of a clear sample container and the aid of focused lighting simple field checks can be conducted at any time. It is recommended that the sample is allowed to settle for approximately 5 min prior to conducting observations.

- Water – Turbine oil should be transparent. Check for elevated water presence by visual inspection. At ambient temperatures a wristwatch face, or other readable material, should be easily read through a clear sample container filled with oil. Haziness suggests water concentrations above 300 ppm. Free water forms in turbine oils as the oil cools and the water comes out of solution. Turbine oils at temperatures of 70°F (21°C) and water concentrations above 100 ppm may see free water. Dissolved water will not be visible to the naked eye.
- Viscosity – An inexpensive flo-stick, a plastic board with two channeled oil paths, can be used as a viscosity screening tool. Relative flow rates can be used as a proxy for viscosity and can be evaluated comparing the in-service oil versus new.
- Sediment – Visual sediment may appear in suspension or settle to the container bottom.
- Foam – After vigorous shaking of a clear sample container the building of surface foam should be difficult to generate. Any stable, surface-level foam may be an indication of reduced foam performance and more detailed analysis may be warranted.
- Demulsibility – Equal parts of boiler water and turbine oil, at ambient temperature, can be vigorously shaken together in a clear container. Two distinct phases should readily appear with complete separation within 60 min. Incomplete phase separation suggests reduced demulsibility performance and more detailed analysis may be warranted.
- Air release – After vigorous shaking of a clear container, air bubbles in the body of the oil sample should dissipate within 5 min.

- Color – Unusual and rapid darkening can indicate contamination or excessive degradation.
- Odor – Sour smelling oil can indicate contamination or excessive degradation.

Routine Analysis – Monthly or Quarterly

- Water by ASTM D 6304c *Test Method for Determination of Water in Petroleum Products, Lubrication Oils, and Additives by Coulometric Karl Fischer* or similar – Total water content, dissolved and free water measured in ppm, is critical to turbine oil performance. Water concentrations above 100 ppm at approximately 70°F (21°C) ambient temperature often exceed the oil's saturation and become "free" water. As the oil temperature increases in operation, the free water will re-solubilize. The water saturation point at 100°F (40°C) is approximately 200 ppm. Free water in steam turbine lubrication systems leads to reduced bearing life through rust, sludge, and the potential for microbial growth. Oils maintained relatively dry, at or below 100 ppm water, offer optimum turbine oil performance and oil life.
- ASTM D 5185 *Test Method for the Determination of Additive Elements, Wear Metals, and Contamination in Used Lubricating Oils and Determination of Selected Elements in Base Oils by Inductively-Coupled Plasma Atomic Emission Spectrometry* – Most non-gear turbine oils will test to yield under 5 ppm in all metals tested in analysis.
 - Phosphorus and or zinc may be present in turbine oils with gear loading performance capabilities.
 - Tin concentrations may be attributed to the tin-babbitted material in the main journal bearings
 - Phosphorus levels above new oil in the absence of other contaminant metals would likely indicate a contamination of phosphate ester (EHC fluid). In cases of phosphate ester contamination, concentrations above 5% (50,000 ppm) are considered excessive as the phosphate ester may soften paints and elastomers.
 - Copper can be the result of tube wear in shell and tube heat exchangers.
 - Be watchful for engine oil contaminants, often less than 5 ppm, which can drastically impact demulsibility. Levels of calcium, phosphorus, zinc, and magnesium suggest engine oil contamination. Engine oil concentrations at 300 ppm (0.03%), less than 2 gal (7.6 l) in a 6,000 gal (22,700 l) tanker have been determined to impact turbine oil demulsibility.
- ASTM D 664 *Test Method for Acid Number of Petroleum Products by Potentiometric Titration* – A sharp increase of 0.3–0.4 mg/KOH over the oil's initial value may indicate contamination or excessive degradation. Note: ASTM D 664 (potentiometric method) is preferred over ASTM D 974 (colorimetric method) for in-service analysis, as darker in-service samples can interfere with test accuracy.
- ASTM D 445 *Test Method for Kinematic Viscosity of Transparent and Opaque Liquids* – An oil viscosity, reported as centistokes at 40°C, increase or decrease by more than 5% of the initial oil's viscosity may indicate contamination or excessive degradation. Changes in viscosity can impact rotor position, radially and axially, which can cause premature wear.
- ISO 4406 *Hydraulic Fluid Power Fluids – Method for Coding the Level of Contamination by Solid Particles* – A typical steam turbine ISO cleanliness target is 18/16/13 or NAS 1638 class 7. Steam turbine users with common hydraulic and turbine oil reservoirs may opt for cleaner targets of 16/14/11 or NAS 1638 Class 5. Elevated particle counts can lead to abrasive wear and potential hydraulic valve sticking.

Advanced Suitability for Continued Use – Annual/Bi-Annual

The suitability for continued use test slate for steam turbine oils should encompass all testing outlined in the routine test slate plus the additional tests outlined below:

- Rotating Pressure Vessel Oxidation Test (RPVOT) (ASTM D 2272) – Considering the typical large volume of oil in a steam turbine oil reservoir and the expense of a full reservoir change-out, turbine oils are formulated to provide high resistance to oxidation. RPVOT testing offers an indication of an oil's remaining life by accelerating the oxidation process through the introduction of heat, water, and metal catalysts. A decrease in RPVOT of the in-service oil of 25% of the oil's initial RPVOT value, with an increase in acid number, is often considered a warning limit. Mixed aromatic amines and hindered phenols antioxidant additive chemistries, as well as those with aromatic amine-only chemistries, do not tend to decay in RPVOT testing as predictably as phenol-only antioxidant chemistries. RPVOT has not been proven to correlate well with actual field formations of sludge or varnish.
- ASTM D 6971 *Test Method for the Determination of Hindered Phenols and Aromatic Amines in*

Non-Zinc Containing Turbine Oils by Linear Sweep Voltammetry – Voltammetry can compare in-service aromatic amine and hindered phenol relative to new oil antioxidant levels in turbine oils. Peak voltage differential values relate to antioxidant levels. Correlations of voltammetric identified antioxidant levels to field deposits are being studied.

- ASTM D 1401 *Test Method for Water Separability of Petroleum Oils and Synthetic Fluids* – The ability of a turbine oil to shed water can be evaluated in ASTM D 1401 testing. New turbine oils typically test to achieve 37 ml of water in 30 min from two 40 ml volumes of sample oil and distilled water. In-service oil is often afforded less stringent guidelines, such as 60 min to achieve 37 ml water. In-service oils may have poor demulsibility performance but can still maintain acceptably low water concentrations with the aid of external water separators or minimal steam gland seal leakage.
- ASTM D 665 A *Test Method for Rust-Preventing Characteristics of Inhibited Oils in the Presence of Water* – Reduced rust protection performance can be confirmed in testing that compares rust formation on a steel test spindle to a test standard coupon. Results reported as “Light Fail” may warrant visual inspection of metal surfaces for signs of rust.
- ASTM D 892 *Test Method for Foaming Characteristics of Lubricating Oils* – If necessary, foam testing should be conducted to confirm a witnessed excessive foam issue in the turbine oil reservoir. Foam levels in a graduated cylinder are measured at three test temperatures, called sequences 1 @ 75°F (24°C), 2 @ 200°F (93.5°C), and 3 returned to 75°F (24°C), and at two conditions, tendency and stability. The tendency measurement is taken immediately after a 5-min air blowing period. The stability measurement is taken after a 10-min settling period. The stability measurement is much more critical, as it indicates an oil’s ability to collapse air bubbles. Stability measurements greater than 0 ml foam should provoke additional investigation.
- Varnish or Sludge Predictors – Currently, there is no standard method for detecting the onset of deposit formation. However two approaches in use with reasonable correlation to field deposits are ultra centrifuge (UC), and membrane patch colorimetry (MPC). Steam turbines may not benefit from varnish or sludge predictors as much as gas turbines with common hydraulic and turbine reservoirs.
 - UC – The sample oils is centrifuged at 17,500 rpm for 30 min and the resulting deposit is rated

against a one to eight scale, eight being the highest deposit rating.

- MPC – The sample oil with diluents is pulled through a specified micron patch and the resulting color is evaluated against reference standards to determine the severity of deposit formation.

Maintenance Strategies

Some experienced steam turbine users opt to employ the testing outlined above as a part of a planned turbine oil maintenance and rejuvenation program, sometimes called “bleed and feed.” Based on periodic testing, a specified volume of steam turbine oil is removed and replaced with new oil. Based on past experience, some steam turbine users simply opt to replace 10% per year. If properly executed this turbine oil maintenance strategy can extend turbine oil life almost indefinitely and minimize the impact of turbine oil performance decay.

Major turbine overhauls provide an opportunity to condition the turbine oil charge during transfer to clean and dry storage tanks. Care must be taken to ensure the cleanliness of the storage tank to prevent microbial growth.

Should a different turbine oil be considered as make-up to the existing charge of oil a compatibility test is recommended. ASTM 7155 – *Standard Practices for Evaluating Compatibility of Mixtures of Turbine Lubricating Oils* offers conversion guidance. Testing described in this standard practice observes changes in blended samples versus neat samples in visual inspection and selected performance tests.

Key Applications

Steam Turbines, Hydrogen seal oil systems, journal bearings, lube oil reservoir, oil analysis, water removal systems

Conclusions

Selection of a high-quality turbine oil that offers superior performance in steam turbine service is the first step in reliable turbine operation. Knowledge of steam turbine system capabilities and limitations also supports reliable operation. And lastly, a well-thought-out maintenance program that includes a comprehensive oil analysis program and oil conditioning is a key element to steam turbine reliability.

Cross-References

- [Elastohydrodynamic Lubrication \(EHL\)](#)
- [Gear Lubricants](#)
- [Hydrodynamic Fixed Geometry Thrust Bearings](#)

- [Hydrodynamic Journal Bearing](#)
- [Lubricant Formulation](#)
- [Lubricant Viscosity](#)
- [Mineral Oil Base Fluids](#)
- [Mixed EHL](#)
- [Thrust Bearings in Power Generation](#)

References

- ASTM International, *ASTM Standards D 1401, D 892, D 445, D 664, D 2272, D 4378, D 4304, D 4927, D 5182, D 5185, D 6304, D 665, D 6439, D 6443, D 7155 and D 6971*, West Conshohocken
- C. Baurer, M. Day, Pall Corporation, Water Contamination in Hydraulic and Lube Systems. *Practicing Oil Analysis Magazine*, Sept 2007
- H.P. Bloch, *Practical Lubrication for Industrial Facilities* (Fairmont, Lithburn, 2000)
- A. Osborne, *Modern Marine Engineer's Manual*, 2nd edn. (Cornell Maritime, Centerville, 1980)
- D.M. Pirro, A.A. Wessol, *Lubrication Fundamentals ExxonMobil Lubricants & Specialties*, 2nd edn. (Marcel Dekker, New York, 2001)

Steel Rolling

- [Chemistry of Rolling Lubricants](#)

Steric Force

- [Hydration Force](#)

Sticking

- [Electrostatic Field Effects on Adhesion](#)

Stiction

- [Friction \(Concepts\)](#)

Stiction in Magnetic Recording

- [Surface Texture for Magnetic Recording](#)

Stochastic Contact Theories: Other Theories Based on the Greenwood-Williamson Model

ROBERT L. JACKSON

Department of Mechanical Engineering, Auburn University, Auburn, AL, USA

Synonyms

[Multiscale contact](#); [Real area of contact](#); [Rough surface contact models](#); [Statistical contact](#)

Definition

This chapter will expand the discussion on the previously introduced statistical models based on the work of Greenwood and Williamson (1966). These models are used to consider the contact of nominally rough surfaces that possess small scale roughness. Especially the inclusion of plastic deformation in these models will be discussed. This is important because isolated asperities often carry very high pressures that can cause yielding. In addition, recent multi-scale models of rough surface contact that differ from the fractal based models will be discussed. These models predict the same general trend as the stochastic or statistical models, but sometimes make quantitatively different predictions. However, they are important in understanding the multiscale nature of surface contact in addition to the statistical nature.

Scientific Fundamentals

Elastic–Plastic Statistical Models

A landmark discovery in contact mechanics and tribology was made when Archard (1957) and then Greenwood and Williamson (1966) found theoretically that the real area of contact and force between contacting rough surfaces have a linear relationship.

As mentioned previously, Greenwood and Williamson (GW) (1966) show that rough surfaces can be modeled as a collection of individual asperities of various heights. These asperities are then categorized by a few statistical parameters describing the surface. First, the GW model assumes that all asperities have the same radius of curvature, R . Then, the distance between the surfaces can be described in two ways: (1) the distance between the mean of the surface heights, h , and (2) the distance between the mean of the summit heights or asperity peaks, d . These values are related by

$$h = d + y_s \quad (1)$$

The value of y_s is derived by Etsion and Front (1994) and given as:

$$y_s = \frac{0.045944}{\eta R} \quad (2)$$

where η is the area density of the asperities.

When the surfaces are pressed together, some of the asperities will interfere a distance ω with the opposing surface. Since the surfaces cannot penetrate each other, ω is also the distance each asperity compresses perpendicular to the surfaces (sometimes referred to as interference). The interference is defined as:

$$\omega = z - d \quad (3)$$

where the height of each asperity is defined by a distance, z , from the mean asperity height. The heights of the asperities are also assumed to have a statistical distribution function, $\phi(z)$. Many times, the uncompromised Gaussian distribution is used, and the integrals are evaluated numerically. Although some past works have used a simplified exponential version of the Gaussian distribution (see Greenwood and Williamson 1966; Etsion and Front 1994). Recently the Greenwood and Williamson model has been solved analytically with the full Gaussian distribution (Jackson and Green 2011).

Then, the total area of contact and total contact force between the surfaces is found by simple integrals over the entire range of asperity contact:

$$A(d) = \eta A_n \int_d^{\infty} \bar{A}(z - d) \phi(z) dz \quad (4)$$

$$P(d) = \eta A_n \int_d^{\infty} \bar{P}(z - d) \phi(z) dz \quad (5)$$

The GW model then assumes that the asperities deform elastically and are defined by the Hertz elastic solution. Detailed descriptions of the Hertz elastic solution are found in most contact mechanics and tribology texts.

Instead of the Hertzian elastic solution, models that account for elasto-plastic deformation of an asperity can be used in (4) and (5). A representation of these elasto-plastic models is outlined below. Because eventually any contact model accumulates statistically the contribution of all asperity contact points, the integration process tends to diminish the deviations between the various models (suggesting dominance by the statistics rather than by the models).

Chang et al. (1987) developed an elastic-plastic contact model (CEB) that supplemented the Greenwood and Williamson elastic contact model. First, the CEB model approximated elasto-plastic contact by modeling a plastically deformed portion of a hemisphere using volume conservation. The CEB model assumptions are (1) that the hemisphere deformation is localized to near its tip, (2) the hemisphere behaves elastically below the critical interference, ω_c , and fully plastically above that value (despite what is claimed in the paper), and (3) the volume of the plastically deformed hemisphere is conserved. Using these assumptions the following approximations for contact area and force in the elastic-plastic range ($\omega/\omega_c > 1$) are analytically derived as

$$\bar{A}_{CEB} = \pi R \omega (2 - \omega_c/\omega) = \pi R (2\omega - \omega_c) \quad (6)$$

$$\bar{P}_{CEB} = \pi R \omega (2 - \omega_c/\omega) KH = \pi R (2\omega - \omega_c) KH \quad (7)$$

where K is the hardness factor given by $K = 0.454 + 0.41\nu$. Also, the critical interference used in the CEB model is given by:

$$\omega_c = \left(\frac{\pi KH}{2E'} \right)^2 R \quad (8)$$

where

$$\frac{1}{E'} = \frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \quad (9)$$

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \quad (10)$$

$E_1, \nu_1, R_1, E_2, \nu_2, R_2$, are the elastic properties and radii of sphere 1 and 2, respectively, and the hardness is assumed to be $H = 2.8 \cdot S_y$. An alternative equation based directly on the yield strength is provided later in (17). In addition, the CEB model is limited to this fixed relationship between the hardness and the yield strength, and the model also contains a discontinuity at ω_c . Using Hertz contact and the CEB asperity model, Green (2002) provided an approximate analytical solution to the rough surface statistical model considering the full Gaussian distribution. Despite the shortcomings mentioned above, the CEB model has been used for many years in many subsequent works for different applications with some success. However, the past decade has brought more advances through the use of finite element modeling.

If the plastic deformation covers the entire area of contact, it is said that a fully plastic condition is reached. The fully plastic truncation model states that under fully plastic conditions the area of contact of an asperity pressed against a rigid flat can be approximately calculated by

truncating the asperity tips as the rigid flat translates an interference, ω . For a hemisphere, this approximated fully plastic area is given by:

$$\bar{A}_P = 2\pi R\omega \quad (11)$$

which predicts larger contact areas than Hertz elastic contact. Using (11), the contact force of the hemispherical asperity is simply the contact area multiplied by the average contact pressure, which in this case is the hardness, since the contact is assumed to be fully plastic. The fully plastic contact force is thus:

$$\bar{P}_P = 2\pi R\omega H \quad (12)$$

Since plastic deformation of the asperity will increase the area of contact, the truncation model produces the proper trend to some degree. However, FEM results show that this model is unjustifiable. Although this model is often attributed to Abbott and Firestone (1933), they intended their model to be used to describe a wear process rather than an indentation process.

Kogut and Etsion (2002) performed a FEM analysis of the case of an elastic-perfectly plastic sphere in contact with a rigid flat. Their work gives a very detailed analysis of the stress distribution in the contact region, and empirical expressions are provided for the contact area and the contact force. These are given in a piece-wise form as:

For $1 \leq \omega/\omega_c \leq 6$

$$\bar{P}_{KE} = \bar{P}_c \cdot 1.03 \left(\frac{\omega}{\omega_c} \right)^{1.425} \quad (13)$$

$$\bar{A}_{KE} = \bar{A}_c \cdot 0.93 \left(\frac{\omega}{\omega_c} \right)^{1.136} \quad (14)$$

For $6 \leq \omega/\omega_c \leq 110$

$$\bar{P}_{KE} = \bar{P}_c \cdot 1.40 \left(\frac{\omega}{\omega_c} \right)^{1.263} \quad (15)$$

$$\bar{A}_{KE} = \bar{A}_c \cdot 0.94 \left(\frac{\omega}{\omega_c} \right)^{1.146} \quad (16)$$

These equations are discontinuous at $\omega/\omega_c = 1$ and $\omega/\omega_c = 6$, and they describe the deformation only up to $\omega/\omega_c = 110$, at which point the fully plastic truncation model is assumed. The KE model also assumes the value of H to be fixed at $2.8S_y$. However, these are simple equations that can be useful when easy analytical manipulation is desired. Kogut and Etsion then incorporated their model in the statistical methodology set forth by Greenwood and Williamson to provide an elastic-plastic model of rough surface contact.

Later, Jackson and Green (2005) derived the critical interference in relation to the yield strength rather than

the hardness by using the von Mises yield criterion (VM). The resulting equation is:

$$\omega_c = \left(\frac{\pi \cdot C \cdot S_y}{2E'} \right)^2 R \quad (17)$$

where C is

$$C = 1.295 \exp(0.736\nu) \quad (18)$$

The Poisson's ratio, ν , to be used in (18) is that of the material that yields first.

The critical force, \bar{P}_c , is then calculated from the critical interference, ω_c , by substituting (17) into the Hertz contact solutions. Overbars are used to denote the case of a single asperity model rather than a multiple asperity or surface model. The resulting critical contact force at initial yielding is:

$$\bar{P}_c = \frac{4}{3} \left(\frac{R}{E'} \right)^2 \left(\frac{C}{2} \pi \cdot S_y \right)^3 \quad (19)$$

Similarly, the critical contact area is given by:

$$\bar{A}_c = \pi^3 \left(\frac{CS_y R}{2E'} \right)^2 \quad (20)$$

Jackson and Green (JG) also used the finite element method to predict the contact force and area between an elastic perfectly plastic hemisphere and a flat. In their work a finite element analysis is performed that produced results appreciably different than the similar Kogut and Etsion (KE) model. The model accounts for geometry and material effects that are not accounted for in the KE model. Most notable of these effects is that the predicted geometrical hardness, defined as the uniform pressure found during fully yielded contact, is not constant and changes with the evolving contact geometry and material properties. At $0 \leq \omega/\omega_c \leq 1.9$ the JG single asperity model effectively coincides with the Hertzian solution. This is only an effective description and yielding is still predicted to first occur when $\omega = \omega_c$. At interferences larger than this the following equations describing elasto-plastic single asperity contact are used:

For $\omega \geq 1.9\omega_c$

$$\bar{A}_{JG} = \pi R \omega \left(\frac{\omega}{1.9\omega_c} \right)^B \quad (21)$$

$$\begin{aligned} \bar{P}_{JG} = \bar{P}_c \left\{ \left[\exp \left(-\frac{1}{4} \left(\frac{\omega}{\omega_c} \right)^{\frac{5}{12}} \right) \right] \left(\frac{\omega}{\omega_c} \right)^{3/2} \right. \\ \left. + \frac{4H_G}{CS_y} \left[1 - \exp \left(-\frac{1}{25} \left(\frac{\omega}{\omega_c} \right)^{\frac{5}{9}} \right) \right] \frac{\omega}{\omega_c} \right\} \end{aligned} \quad (22)$$

where

$$B = 0.14 \exp(23 \cdot e_y) \quad (23)$$

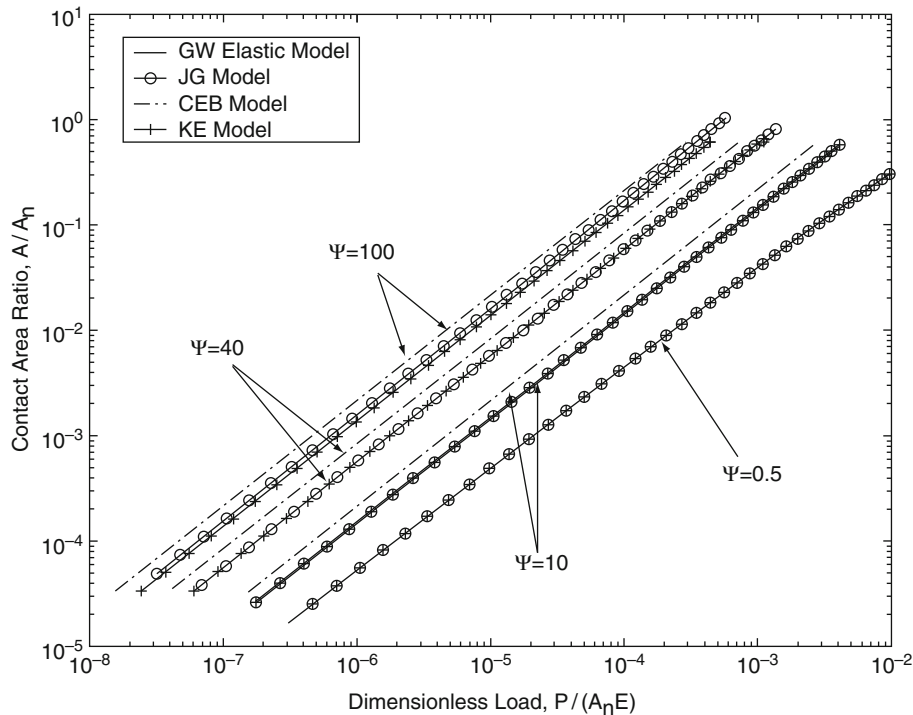
$$e_y = \frac{S_y}{E'} \quad (24)$$

$$\frac{H_G}{S_y} = 2.84 - 0.92 \left[1 - \cos \left(\pi \sqrt{\frac{\omega}{R}} \left(\frac{\omega}{1.9\omega_c} \right)^{B/2} \right) \right] \quad (25)$$

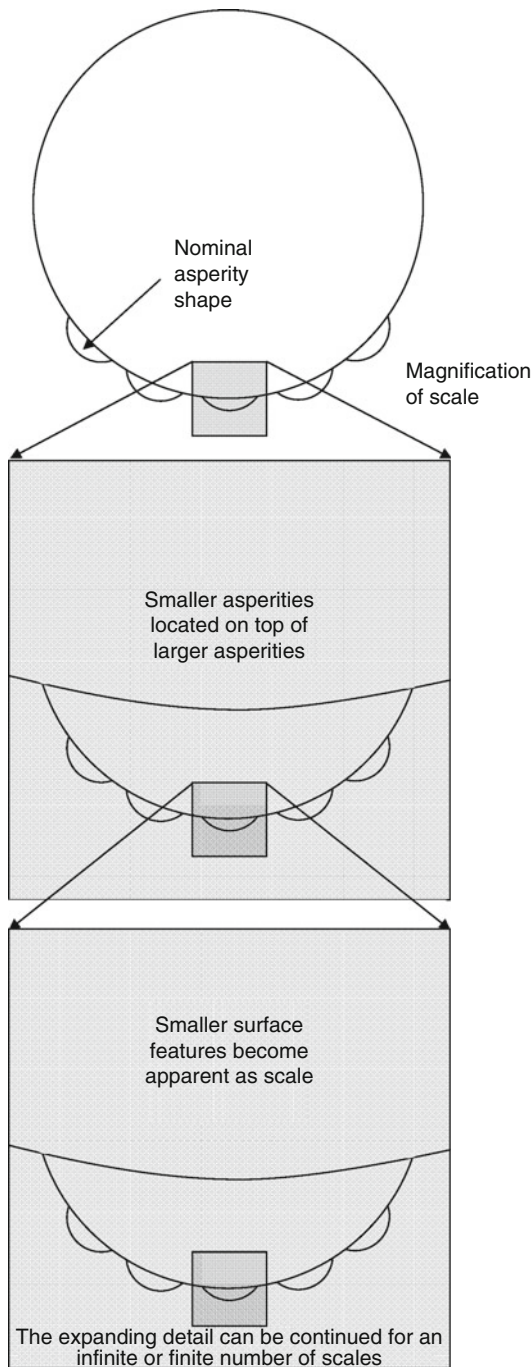
For an arbitrary surface, the predictions of real contact area as a function of load using the Greenwood and Williamson model (GW), the Chang Etsion and Bogy Model (CEB), the Kogut and Etsion Model (KE), and the Jackson and Green model (JG) are compared in Fig. 1. The plasticity index, Ψ , is also varied. One can see that for low plasticity indices, usually indicative of elastic contact, all the models agree very well. However, as more plasticity is included and the plasticity index climbs, the models diverge more and more. Nonetheless, one can see that all of the elastic-plastic models make similar qualitative predictions.

Multiscale Modeling of Rough Surface Contact

An early appreciation for the influence of multiple scales was demonstrated by Archard (1957), whose work actually precedes the GW model. Archard suggested that the asperities of rough surfaces must be modeled as “protuberance upon protuberance.” In other words, any asperity has upon it a collection of smaller asperities, each of which supports a collection of even smaller asperities, which will continue to consider an infinite number of scales, or practically some limiting scale (such as the atomic scale). With some simplifying assumptions, Archard was able to model this scaling effect for the elastic contact between (1) a rough sphere and a smooth, rigid flat and (2) a nominally flat, rough surface and a smooth, rigid flat (see Fig. 2). In the latter case, the rough surface was composed of rough hemispherical asperities, having a uniform height distribution. In both cases, Archard showed that as higher and higher levels of refinement are considered (each of which has a characteristic asperity radius of curvature and asperity areal density), the relationship between area of contact and load approached linearity.



Stochastic Contact Theories: Other Theories Based on the Greenwood-Williamson Model, Fig. 1 A comparison of several statistically based contact models, excerpted from Jackson and Green (2006) (Used with permission from Elsevier B. V.)



Stochastic Contact Theories: Other Theories Based on the Greenwood-Williamson Model, Fig. 2 Schematic showing Archard's concept of multiple scales of stacked surface features or asperities

However, Archard's contact model is based on hypothetical, idealized surfaces and is difficult to apply to a real rough surface. For example, Archard's model does not provide a means of determining the required coefficients from measurements of a surface profile. Archard assumed that each successive level had asperity radii much smaller than in the previous level. Because real surfaces would generally not have successive scales so widely separated, it is problematic to identify such in practice.

One of the first multiscale methodologies that set out to expand and implement Archard's concept is by Ciavarella et al. (2000), which models contact using an idea similar to Archard's and the current work. They modeled the surface structure by using the popular Weierstrass–Mandelbrot (W–M) fractal equation and also use a two-dimensional elastic sinusoidal model (Westergaard 1939). They then used the Archard concept and assumed that the scales given by the W–M fractal were stacked upon each other. They also conclude that as higher scales are included in the contact model via fractal mathematics that the contact area will approach zero. This is a result of assuming that the surface is characterized by the W–M fractal.

Jackson and Streator presented another method using the same direction of thought as Archard (1957), but provides a method that can be much more easily applied to real surfaces. The model assumptions, which are somewhat different from those in the GW and MB fractal models, are as follows:

1. Asperities are arranged so that asperities of smaller cross-sectional surface area are located on top of larger asperities. In the frequency domain this means that asperity distributions of higher frequencies are superimposed upon lower frequency asperities. This is similar to Archard's "protuberance upon protuberance" concept.
2. Each "level" or frequency of asperities carries the same total load.
3. The load at each frequency level is shared equally among all the asperities at that level.
4. At a given frequency level, each asperity deforms according to Hertz theory or to a chosen elasto-plastic asperity contact model, irrespective of the presence of higher frequency asperities upon it.
5. A given frequency level cannot increase the contact area beyond what is experienced by the frequency level below it.

These assumptions set up the following simple framework of equations for the contact model:

$$A_r = \left(\prod_{i=1}^{i_{\max}} \bar{A}_i \eta_i \right) A_n \quad (26)$$

$$F = \bar{F}_i \eta_i A_{i-1} \quad (27)$$

where A_r is the real area of contact, F is the contact load, A_n is the nominal contact area, and the subscript i denotes a frequency level, with i_{\max} denoting the highest frequency level considered. Parameters \bar{A}_i and \bar{F}_i are the single asperity contact area and single asperity contact force at a given frequency level, respectively. The total (nominal) area of contact at a given frequency level is denoted by A_p , while η_i is the corresponding areal asperity density. For example, if a simplified hypothetical case is assumed such that there are only two frequency levels of asperities then (1) becomes

$$A_r = \bar{A}_2 \eta_2 \bar{A}_1 \eta_1 A_n \quad (28)$$

where \bar{A}_2 and \bar{A}_1 are the single asperity contact areas and η_1 and η_2 are the areal asperity densities at their respective frequency levels. The product of $\eta_1 A_n$ gives the number of asperities at frequency level 1. Thus the value of $\bar{A}_1 \eta_1 A_n$ gives the contact area of frequency level 1, which is viewed as the nominal contact area from the perspective of level 2. Next, the value of $\eta_2 \bar{A}_1 \eta_1 A_n$ gives the number of asperities at frequency level 2, which, when multiplied by \bar{A}_2 , the contact area per level 2 asperity, yields the real contact area, A_r . Values for \bar{A}_1 and \bar{A}_2 are determined from a micro-contact model (e.g., Hertz), assuming that the contact load is equally shared by all asperities of a given level, with the asperity radius of curvature established from the frequency spectrum. In the general case, the repetitive cycle continues until all the asperity frequency levels are considered. Thus, the resulting model uses a recursive

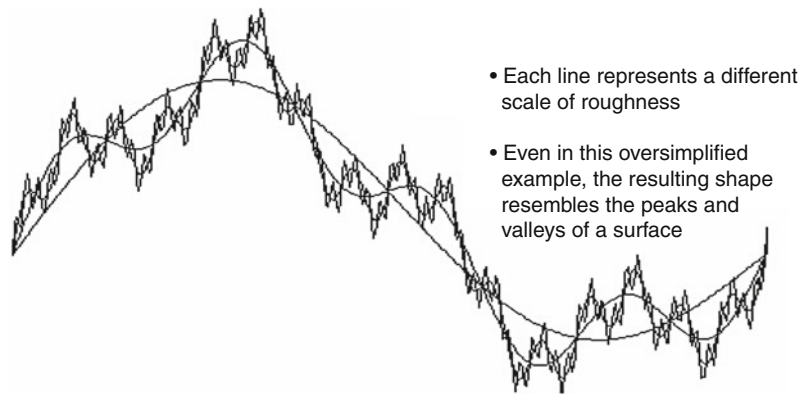
approach to predict the real area of contact between two rough surfaces.

A variety of individual asperity models are available for use within the multi-scale framework and the other rough surface contact models described above to relate the contact area to the contact force. For a simulation of purely elastic rough surface contact, the Hertzian model may be applied. Alternatively, the several models that were previously discussed that account for elasto-plastic deformation of spherically shaped asperities can also be used.

Since the multiscale models assume that the surfaces are composed of a series of sinusoidal shaped asperities, it is logical to use sinusoidal contact models (see Fig. 3). Considering again the case of elastic contact, the model framework can also employ the solution to the problem of a sinusoidal surface in contact with a flat, which was solved first by Westergaard (1939) for the 1D waviness case. The analysis for the case of 2D waviness developed by authors Johnson, Greenwood, and Higginson provides a relation between pressure and contact area (Johnson et al. 1985). First let \bar{p} be defined as the average pressure in the interface (considering both contacting and non-contacting regions) and let p^* be the amplitude of a sinusoidal pressure superimposed on the mean pressure. Special relations arise when the pressure amplitude p^* is given by:

$$p^* = \sqrt{2\pi E' \Delta f} \quad (29)$$

where E' is the reduced elastic modulus, β is the waviness amplitude, and f is the reciprocal of wavelength. Namely, when $\bar{p} \geq p^*$, the pressure loads the surfaces so that there is no gap between them. For such a case, the area of contact for the corresponding frequency level is taken to equal that of the frequency level below it, so that the asperities at the



Stochastic Contact Theories: Other Theories Based on the Greenwood-Williamson Model, Fig. 3 Depiction of a rough surface consisting of a series of superimposed sinusoidal waves, excerpted from Wilson et al. (2010) (Used with permission from Elsevier B. V.)

given frequency level induce no separation between the surfaces. Alternatively, when $\bar{p} < p^*$ the contact is not complete, and a closed form solution for the two-dimensional waviness contact problem is not available.

If the surfaces are sufficiently loaded, they will yield and undergo plastic deformation. For elastic-plastic sinusoidal contact the following equation is used to predict the pressure to cause complete contact (Jackson et al. 2008a):

$$\frac{p_{ep}^*}{p^*} = \left(\frac{11}{4\Delta/\Delta_c + 7} \right)^{3/5} \quad (30)$$

Where Δ_c is the critical amplitude. When $\Delta < \Delta_c$ plastic deformation of the sinusoidal surface never occurs over the entire contact regime (from initial to complete contact). Δ_c is analytically calculated using the von Mises yield criteria and the stress functions provided by (Tripp et al. 2003) and is given by:

$$\Delta_c = \frac{\sqrt{2}}{3\pi} \frac{S_y}{E'} \frac{e^{\frac{2}{3}\nu}}{f} \quad (31)$$

Using the multiscale technique, the area will continue to decrease as scales are iteratively included until the asperities begin to flatten out because the real contact pressure at each frequency level is greater than p^* or p_{ep}^* . If the contact pressure is less than these values at a particular scale than that scale will reduce the area of contact and the contact pressure will increase. Again, this will continue until the contact pressure overcomes the p^* or p_{ep}^* . Therefore, for fractal contact using this stacked asperity concept the real contact pressure approaches p^* or p_{ep}^* . The real area of contact between the surfaces is then given by (Jackson et al. 2008b):

$$A_r = \frac{F}{p^*} \quad (32)$$

First, if a surface is assumed to be described by the Weierstrass-Mandelbrot function, Eq. 32 can be solved using the W-M fractal parameters (Ciavarella et al. 2000). Now define B as the average ratio of the amplitude (Δ) of the sinusoidal surface at each scale to the wavelength (λ) of that scale. Since γ^n is related to the inverse of the wavelength, B is related to the fractal description by

$$B = \left(\frac{\Delta}{\lambda} \right) = \frac{G^{(D-1)}\gamma^n}{\gamma^{(2D-1)n}} = \frac{G^{(D-1)}}{\gamma^{(1-D)}} = G^{(D-1)}\gamma^{(D-1)n} \quad (33)$$

This results in B increasing with smaller scales (n). Unfortunately this means that the pressure required to obtain complete contact at each scale will then increase with smaller scales, as given by (29) and (30). This results in the real area of contact reducing with each included

scale, until approaching zero when an infinite number of scales are included. This was also shown more rigorously by Ciavarella et al. (2000).

Alternatively, by using the same methodology but assuming that the ratio of amplitude to wavelength (B) is constant over all scales the following equations arise (Jackson et al. 2008a):

$$(A_r)_{\text{elastic}} \approx \frac{F}{\sqrt{2\pi E' B}} = \frac{F}{P_r} \quad (34)$$

$$(A_r)_{\text{elastic-plastic}} \approx \frac{F}{\sqrt{2\pi E' B}} \left(\frac{\frac{12\pi E'}{\sqrt{2} S_y e^{\frac{2}{3}\nu}} B + 7}{11} \right)^{3/5} = \frac{F}{P_r} \quad (35)$$

For surfaces that do not possess a constant B over all scales, an approximate solution can be found by using the maximum value of B over all the considered scales. In addition, since B is assumed to describe the shape of every roughness scale on the surface, it can be related to the RMS slope (which can be related to the real contact pressure as is done by several other modeling methodologies (Pei et al. 2005; Whitehouse and Archard 1970)). The average slope of a sine wave which in the current case is also the RMS slope is then easily calculated to give

$$\bar{m} = \sqrt{2\pi B} \quad (36)$$

where \bar{m} is the RMS slope. Note that (36) is only valid if B is constant over all scales. Then by substituting (36) into (34), the elastic multiscale model is given as

$$(A_r)_{\text{elastic}} = \frac{F}{E' \bar{m}} \quad (37)$$

Therefore the above solutions can also be correlated with the models predicted by Bush et al. (1975) and Persson (2001) which are given, respectively, as:

$$(A_r)_{\text{elastic}} = \sqrt{2\pi} \frac{F}{E' \bar{m}} \quad (38)$$

$$(A_r)_{\text{elastic}} = \sqrt{\frac{8}{\pi}} \frac{F}{E' \bar{m}} \quad (39)$$

Bush et al. (1975) derive their model similar to a Greenwood and Williamson type statistical model, while Persson (2001) uses a diffusion theory to model rough surface contact. Interestingly, all models are within an order of magnitude agreement. Hyun et al. (2004) and Pei et al. (2005) also show that the models by

Bush et al. (1975) and Persson (2001) appear to agree fairly well with deterministic finite element models. Jackson and Green (2011) also showed that Eq. (37) agrees very well with deterministic solutions of elastic rough surface contact solved in the frequency domain.

Key Applications

Prediction of friction and stresses between mechanical components such as gears, bearings, cams, bolts, MEMS, etc.

Prediction of electrical and thermal contact resistance in contacts.

Cross-References

- ▶ [Adhesive Contact of Elastic Bodies](#)
- ▶ [Adhesive Contact of Inelastic Bodies](#)
- ▶ [Contact of Rough Surfaces: The Greenwood and Williamson/Tripp, Fuller and Tabor Theories](#)
- ▶ [Contact Yield](#)
- ▶ [FFT-Based Methods for Contact Mechanics](#)
- ▶ [Finite Element Approach for Contact Simulation](#)
- ▶ [Fractal Contact Mechanics](#)
- ▶ [Numerical Methods for Elastic Contact Problems](#)
- ▶ [Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics](#)

References

- E.J. Abbott, F.A. Firestone, Specifying surface quality—a method based on accurate measurement and comparison. *Mech. Engr.* **55**, 569–572 (1933)
- J.F. Archard, Elastic deformation and the laws of friction. *Proc. R. Soc. Lond. A* **243**, 190–205 (1957)
- A.W. Bush, R.D. Gibson et al., The elastic contact of rough surfaces. *Wear* **35**, 87–111 (1975)
- W.R. Chang, I. Etsion et al., An elastic-plastic model for the contact of rough surfaces. *ASME J. Tribol.* **109**(2), 257–263 (1987)
- M. Ciavarella, G. Demelio et al., Linear elastic contact of the Weierstrass profile. *Proc. R. Soc. Lond. A* (456), 387–405 (2000)
- I. Etsion, I. Front, Model for static sealing performances of end face seals. *Tribol. Trans.* **37**(1), 111–119 (1994)
- I. Green, A transient dynamic analysis of mechanical seals including asperity contact and face deformation. *Tribol. Trans.* **45**(3), 284–293 (2002)
- J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces. *Proc. R. Soc. Lond. A* **295**, 300–319 (1966)
- S. Hyun, L. Pel et al., Finite-element analysis of contact between elastic self-affine surfaces. *Phys. Rev. E* **70**(2.2), 026117-1–026117-12 (2004)
- R.L. Jackson, I. Green, A finite element study of elasto-plastic hemispherical contact. *ASME J. Tribol.* **127**(2), 343–354 (2005)
- R.L. Jackson, I. Green, A statistical model of elasto-plastic asperity contact of rough surfaces. *Tribol. Int.* **39**(9), 906–914 (2006)
- R.L. Jackson, I. Green, On the modeling of elastic contact between rough surfaces. *Tribol. Trans.* **54**(2), 300–314 (2011)
- R.L. Jackson, V. Krithivasan et al., The pressure to cause complete contact between elastic plastic sinusoidal surfaces. *IMechE Part J: J. Eng. Tribol.* **222**(7), 857–864 (2008a)

- R.L. Jackson, W.E. Wilson et al., in *A Study of the Average Real Contact Pressure between Rough Surfaces*. 2008 STLE/ASME International Joint Tribology Conference, Miami, FL, 2008
- K.L. Johnson, J.A. Greenwood, J.G. Higginson, The contact of elastic regular wavy surfaces. *Int. J. Mech. Sci.* **27**(6), 383–396 (1985)
- L. Kogut, I. Etsion, Elastic-plastic contact analysis of a sphere and a rigid flat. *J. Appl. Mech. Trans. ASME* **69**(5), 657–662 (2002)
- L. Pei, S. Hyun et al., Finite element modeling of elasto-plastic contact between rough surfaces. *J. Mech. Phys. Solids*. **53**(11), 2385–2409 (2005)
- B.N.J. Persson, Elastoplastic contact between randomly rough surfaces. *Phys. Rev. Lett.* **87**(11), 116101 (2001)
- J.H. Tripp, J.V. Kuilenburg et al., Frequency response functions and rough surface stress analysis. *Tribol. Trans.* **46**(3), 376–382 (2003)
- H.M. Westergaard, Bearing pressures and cracks. *ASME J. Appl. Mech.* **6**, 49–53 (1939)
- D.J. Whitehouse, J.F. Archard, The properties of random surfaces of significance in their contact. *Proc. R. Soc. Lond. A* **316**, 97–121 (1970)
- W.E. Wilson, S.V. Angadi, R.L. Jackson, Surface separation and contact resistance considering sinusoidal elastic-plastic multi-scale rough surface contact. *Wear*, **268**(1–2), 190–201 (2010)

Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics

J. A. GREENWOOD

Department of Engineering, University of Cambridge, Cambridge, UK

Synonyms

[Nayak's theory](#); [Random field theory](#)

Definition

Surface roughness may be treated as a random field, the two-dimensional equivalent of Rice's classic theory of noise down a telephone line. Many relevant properties of the roughness can then be predicted from probability theory.

Scientific Fundamentals

Surface Roughness Theory

The random field theory of Longuet-Higgins (1957a, b) and Nayak (1971) assumes that surface roughness is an assembly of sinusoidal waves in random directions, with a continuous spectrum of frequencies. Longuet-Higgins (1957a) analyzes a general anisotropic surface; Longuet-Higgins (1957b) and Nayak focus on an

isotropic surface. For such a surface the auto-correlation function (ACF) along a profile in any direction is

$R(r) = \lim_{L \rightarrow \infty} \frac{1}{2L} \int_{-L}^{+L} z(x)z(x+r) dx$ and can be estimated directly from height readings taken along the profile, or as the Fourier transform of the profile power spectrum (PSD) $G(k)$:

$$R(r) = \int_0^\infty G(k) \cos(kr) dk;$$

$$G(k) = \frac{2}{\pi} \int_0^\infty R(r) \cos(kr) dr$$

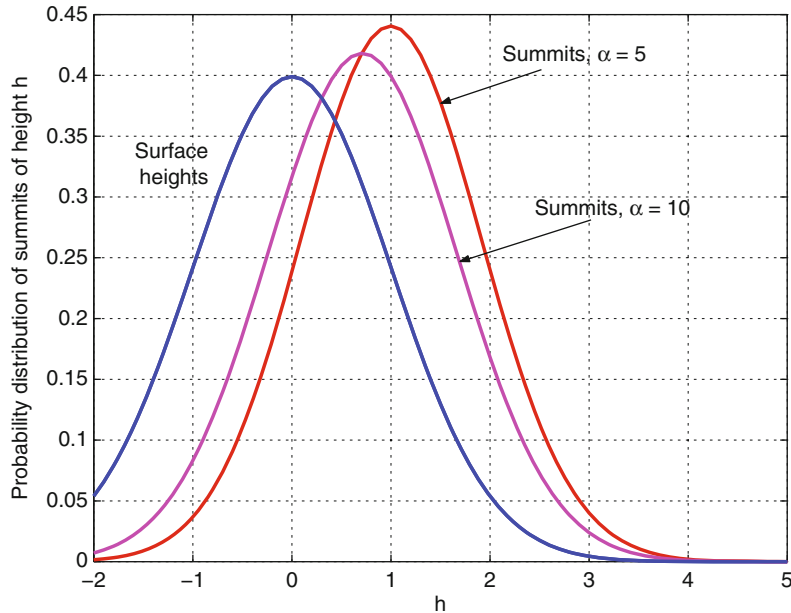
$G(k)$ is more convenient than the surface power spectrum $\Phi^s(k)$, related to the ACF by $\Phi^s(k) = \frac{1}{2\pi} \int_0^\infty R(r) J_0(kr) r dr$: the profile PSD and sur-

face PSDs are related by $G(k) = 4 \int_{t=k}^\infty \Phi^s(t) \frac{t dt}{\sqrt{t^2 - k^2}}$ (Nayak 1973). The two contain the same information, but in a somewhat distorted form: for example, the ACF of a ground surface is close to a simple exponential $R(x) = \sigma^2 \exp(-\beta x)$: the PSD becomes $G(k) = (2/\pi) \beta \sigma^2 / (\beta^2 + k^2)$ while $\Phi^s(k) = \beta \sigma^2 / ((2\pi) (\beta^2 + k^2)^{3/2})$ decreasing more quickly as k increases (shorter wavelengths). More seriously, the PSD is

sometimes “filtered” by imposing a sharp cut-off. Nayak (1973) shows that this implies a singularity in the surface power spectrum, which is impossible. In other words, the PSD should not be truncated.

The features of the roughness of importance in contact will then depend on the mean square roughness $\sigma^2 \equiv m_0 = R(0) = \int_0^\infty G(k) dk$ and the second and fourth moments m_2, m_4 of the profile power spectrum, i.e., $m_n = \int_0^\infty k^n G(k) dk$. [In terms of the surface spectrum, the equations are $m_n = c_n \int_0^\infty k^{n+1} \Phi^s(k) dk$ with $c_0 = 2\pi, c_2 = \pi, c_4 = 3\pi/4$.]

In particular the non-dimensional combination of the three moments $\alpha = m_0 m_4 / m_2^2$, sometimes called the bandwidth parameter, plays a vital role. From the basic assumption in these theories about the nature of surface roughness and the central limit theorem, the *surface* height distribution must be Gaussian (or “Normal”): the height distributions of *peaks* (maxima of a profile) and *summits* (surface maxima) (Fig. 1) will not be Gaussian, although for most values of α the non-gaussianity is slight: however the standard deviations of these distributions differ from that of the surface so that $\sigma_p^2 = \sigma^2 [1 - 0.5708/\alpha]$ and $\sigma_s^2 = \sigma^2 [1 - 0.8968/\alpha]$: note also that the mean planes of the peak and summit height distributions are appreciably



Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics, Fig. 1 Theoretical distribution of summit heights. To the eye, these appear to be Gaussian

above the surface mean plane by amounts $\sqrt{\pi/2} \sigma/\sqrt{\alpha}$ and $(4/\sqrt{\pi}) \sigma/\sqrt{\alpha}$.

Theory also predicts the (linear) density of peaks and zero-crossings of the profile $D_p = (1/2\pi)\sqrt{m_4/m_2}$ and $D_z = (1/\pi)\sqrt{m_2/m_0}$, and the (areal) density of summits $\eta = (1/6\pi\sqrt{3})m_4/m_2$ (so that $\eta \sim 1.2D_p^2$, not $\eta = D_p^2$ as Greenwood and Williamson (GW 1966) assumed).

Nayak's contribution was to predict the distributions of peak and summit curvatures, which are not Gaussian but close to Rayleigh and Γ_2 distributions respectively (see Fig. 2): in particular the mean and standard deviation of the summit curvatures are $(8/3\sqrt{\pi})\sqrt{m_4}$ and $0.518\sqrt{m_4}$ (independent of α).

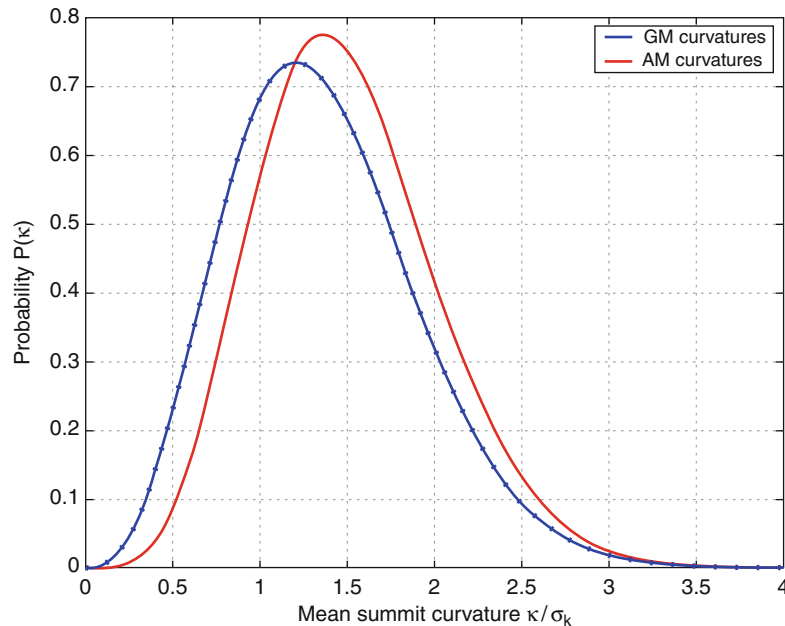
Thus, the input data for the GW theory of rough surface contact can be predicted from the moments m_0 , m_2 , and m_4 . But note also Longuet-Higgins' suggestion that the moments can conveniently be estimated by counting zero-crossing and peak densities (and one can see that for his interest, oceanography, this would be more practical).

According to Nayak's theory, not merely do summits have a range of curvatures (as observed by Greenwood and Williamson, but ignored in the GW theory), but the mean summit curvature varies with height (Fig. 3); the higher summits have much larger curvatures. Additionally, asperities are not paraboloidal but ellipsoidal, although for an

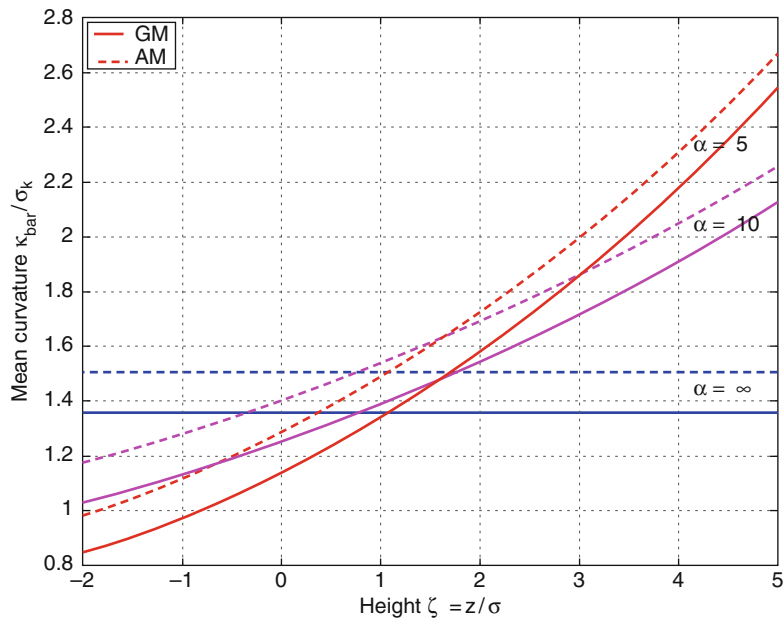
isotropic surface the eccentricity is moderate, the most common ratio of the principal curvatures at a summit being 2:1 (Fig. 4).

The GW contact theory assumes that all summits are axisymmetric (to give circular contacts), all with the same curvature. Bush et al. (1975) attempted the full solution for rough surface contact taking into account the predictions of random field theory: that the summits are ellipsoidal, that the mean curvature increases with the summit height, and that summits of a given height have a distribution of curvatures; an error (see Carbone and Bottiglione 2008) resulted in incorrect numerical values for the real contact pressure, but their discovery that the real pressure is asymptotically constant and equal to $E^* \sqrt{m_2/\pi}$ remains correct (Fig. 5).

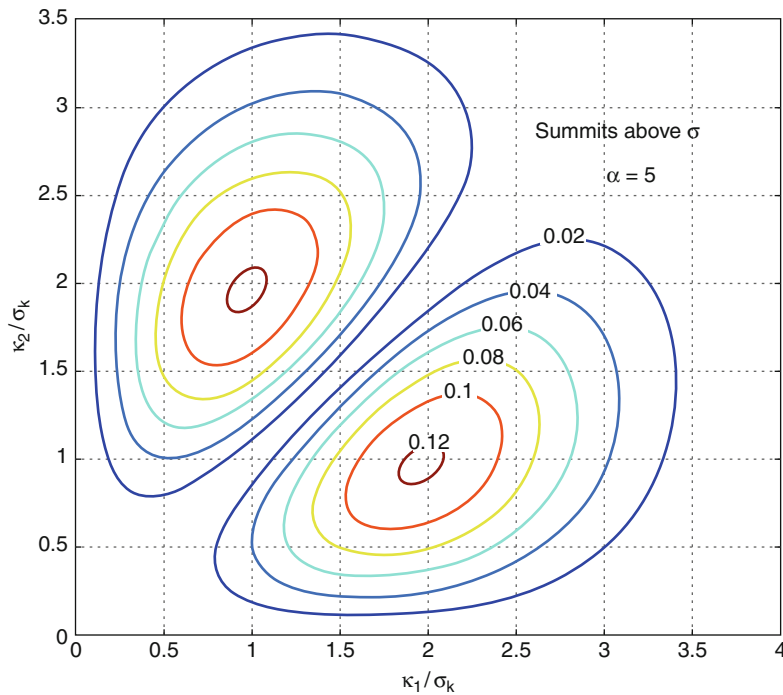
Subsequently (Greenwood 2006; Carbone and Bottiglione 2008) it has become clear that this is primarily the result of the increase of mean summit curvature with height, the elliptical contacts and the scatter in size at a given height being rather unimportant. The contacts, for an isotropic surface, are only mildly elliptical, and this can largely be allowed for by treating them as circular with the GM summit curvature $\sqrt{\kappa_1\kappa_2}$ [in preference to Nayak's AM summit curvature $(\kappa_1 + \kappa_2)/2$].



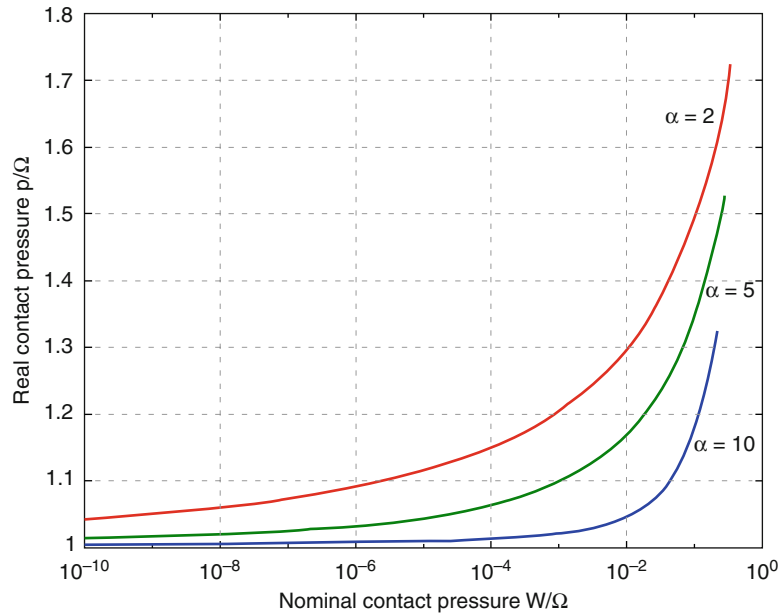
Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics, Fig. 2 Distribution of summit curvatures



Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics, Fig. 3 Variation with height of the mean summit curvature. Both the geometric mean *GM* and arithmetic mean *AM* curvatures are shown. Nayak considers only the *AM*, but the *GM* has advantages (Greenwood 2006)



Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics, Fig. 4 Probability distribution of the principal curvatures at a summit. Summits on an isotropic surface are mildly elliptical, but equal principal curvatures are very rare



Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics, Fig. 5 Variation with load of the mean real pressure. Although the mean real pressure is asymptotically constant, giving direct proportionality between contact area and load, this only occurs at implausibly light loads, where other effects may occur

Effect of a Finite Sampling Interval

As explained above, the PSD (profile power spectrum) $G(k)$ is the Fourier transform of the profile auto-correlation function (ACF) $R(x)$:

$$G(k) = \frac{2}{\pi} \int_0^{\infty} R(x) \cos(kx) dx.$$

Whitehouse and Archard (1970) found that ground surfaces were accurately Gaussian, but that their ACF was a simple exponential decay $R(x) = \sigma^2 \exp(-\beta x)$; accordingly the PSD becomes $G(k) = (2/\pi)\beta \sigma^2/(\beta^2 + k^2)$, and both second and fourth moments m_2, m_4 become infinite; from random field theory it follows that so also do the peak density and peak curvature!

The direct measurement of peak density (or peak curvature) is impractical: experimentally height measurements are made at a sampling interval Δ , and from three consecutive heights z_{-1}, z_0 and z_1 , a peak is defined by $z_0 > z_{-1}, z_1$, and curvature as $\kappa = (z_1 - 2z_0 + z_{-1})/\Delta^2$. Thus, Whitehouse and Archard measured peak density and curvature as a function of the sampling interval Δ and found that these did indeed appear to tend to infinity.

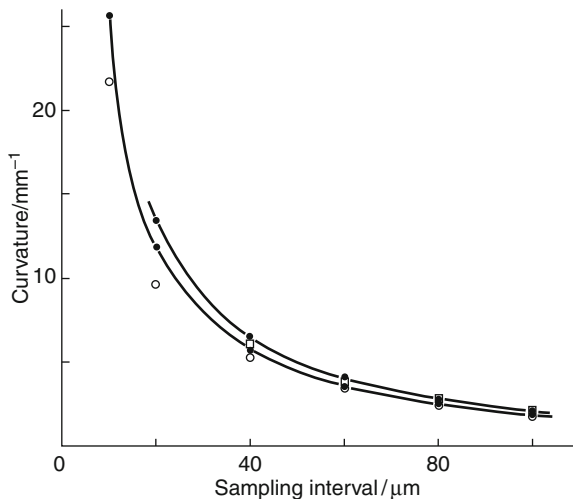
Whitehouse and Archard analyzed the properties of a profile with an exponential ACF (a “Markov process”) as found using a finite sampling interval Δ , and discovered that for such a profile, between one in three and one in

four of all sample points will be a peak! In other words, the peak *density* depends entirely on the sampling interval chosen. Similarly, the mean peak curvature is the product of the rms profile curvature σ_κ with a numerical factor lying between $(2/\sqrt{\pi})$ and $(3\sqrt{3}/2\sqrt{2\pi})$ [1.13–1.04]; but (σ_κ) too is a function of sampling interval and can take any value from zero to $\sigma\sqrt{6}/\Delta^2$ (that is, to infinity if the sampling interval could be reduced to zero!). Measured peak curvature distributions agreed well with theory. Unless an “appropriate” sampling interval can be chosen, the basic parameters for the GW theory are almost arbitrary. [The same problem is there in the BGT theory: with an exponential ACF, the important parameter, the second moment, m_2 , is infinite; replacing it by the mean square profile slope we find $\sigma_m^2(\Delta)$ lies between zero and $2\sigma^2/\Delta^2$.]

A general theory of the properties of a surface as a random field with any ACF as measured at a finite sampling interval has been developed (Whitehouse and Phillips 1982; Greenwood 1984), thus evading the question of the limits as the sampling interval tends to zero but leaving the question of the choice of the “appropriate” sampling interval unanswered. (But see Thomas (1999) for a suggestion that to describe the elastic behavior following initial plastic deformation, the sampling interval should be chosen as the value which makes the plasticity index equal to one.) The surface properties are now found

to depend on two parameters: $r(\Delta)$ and $\theta(\Delta)$. The first is the equivalent of Nayak's α (if $r(0)$ were finite, then $r(0) = 1/\sqrt{\alpha}$). The second is the *sampling-interval* parameter θ defined by $\sin \theta = (\Delta \cdot \sigma_\kappa / 2\sigma_m)$. This would be zero for a Nayak surface (σ_κ , σ_m finite and $\Delta \rightarrow 0$) and must lie between $\pi/4$ and $\pi/3$ for a surface with an exponential ACF: the peak density becomes $\theta/\pi \Delta$ (or the probability that a sample point is a peak becomes θ/π). The mean peak curvature is $\sqrt{\pi/2} \sin \theta/\theta$ while the peak curvature distribution, although dependent on θ , is always close to Nayak's Rayleigh distribution.

The theory does not predict the properties of true summits but only the properties of the "summits" found experimentally by sampling over a uniform square grid, that is, "four-point summits," points higher than their four immediate neighbors. This procedure inevitably locates false summits, an obvious type being a point on a narrow ridge oriented at 45° to the grid of sample points, and it appears that some 30% of the "summits" found will be false. However, the mean curvature of four-point summits when $\theta \rightarrow 0$ is only slightly less than Nayak's answer $1.5045\sigma_\kappa$. The numerical factor is unimportant compared with the real problem brought out by the theory: the significance of the sampling interval. The theory correctly predicts experimentally measured values, but these vary by orders of magnitude depending on the sampling interval chosen (Fig. 6).



Comparison of measured peak, ○, and summit, □, properties for a grit-blasted surface with predicted values, ●. (Data for summit properties at $h = 20 \mu\text{m}$ have not been suppressed but, sadly, lost)

Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics, Fig. 6 Peak and summit curvatures for a grit-blasted surface. Data provided by RS Sayles (see Sayles and Thomas (1979))

Fractal Theory

The immediate reaction in the 1970s to Whitehouse and Archard's measurements and the theoretical consequences was that such behavior was impossible; in the real world curves must have a slope and a finite number of maxima. But Sayles and Thomas (1978) went further and suggested that, throughout nature, real PSDs might be a simple power law $G(k) = A/k^2$ over the whole range of wavenumbers from 0 to ∞ , so that not only were m_2 and m_4 infinite, with unfortunate consequences for slopes and curvatures, but so also was m_0 , striking at the heart of surface roughness specification based on $R_p = \sqrt{m_0}$. Finite values, it was proposed, were obtained only by virtue of inadvertent filtering, with a short wave limit (high k) imposed by a finite stylus size and a long wavelength limit (low k) by a finite sample length or finite specimen size. If these limits are avoided, the profile looks the same whatever magnification is used to observe it; any small part is just as rough as the whole! [The magnifications along and normal to the profile must usually be increased by different factors.] These ideas, and similar considerations of the perimeter of a snowflake or of the coastline of the UK among many others, led Mandelbrot to propose his fractal theory: that natural boundaries are irregular and unmeasurable, and only mathematical idealizations are smooth and differentiable. He noted the classic example from pure mathematics of the Weierstrass function $z(x) = \sum_{n=1}^{\infty} \frac{\cos 2\pi\gamma^n x}{\gamma^{(2-D)n}}$, which although everywhere continuous is nowhere differentiable; that, he argued, is the typical real curve. Note that in comparison with a Fourier series, the wave numbers are in geometrical progression, $1, \gamma, \gamma^2, \dots$ instead of arithmetic progression $1, 2, 3, \dots$: γ is an arbitrary scaling factor, frequently chosen to be 1.5: D is called by Mandelbrot the "fractal dimension" of the curve. The term is justified by the fact that $D = 1$ gives a normal curve that has the proper dimension 1 of a line: but as $D \rightarrow 2$ the curve fills more and more of the space, so may properly be regarded as an area with dimension 2.

Much effort has gone into ingenious ways of generating fractal curves and surfaces, particularly using Mandelbrot's generalization of the Weierstrass function, the W-M function $z(x) = \sum_{n=-\infty}^{\infty} \frac{1 - \cos 2\pi\gamma^n x}{\gamma^{(2-D)n}}$, or by the random mid-point displacement method, where new points are continuously added as the mean of the two (or for a surface, the four) nearest neighbors plus a random addition from a distribution whose standard deviation decreases as a power of the current sampling interval. The results often strongly resemble surface

profiles or surfaces *to the eye*; whether they *behave* like real profiles or surfaces is less clear.

What is clear is that spectral densities are often simple power laws over a very large range of wavenumbers, so that any measurements or physical effects within that range will be dominated by fractal concepts. However much the surface is magnified it will remain rough; contact areas on one scale will prove to be assemblies of smaller contacts at a smaller scale, so predictions of the number or size of contacts are meaningless. The real area of contact or the real contact pressure become meaningless terms.

Contact of Real Surfaces

The summits expected and found on surfaces with power law spectral densities (whether fractal or not) have one saving grace: they are very small, both in height and in spacing. Need they be treated as independent contacts? If an area containing a multitude of summits (a “contour area” in Russian work) is generally in contact (a) do the individual contacts deform into a single contact? (b) does it matter whether they do or not? It is worth noting that in a finite element solution of the contact of a real surface, the load is always applied as point forces at the element nodes, never as a pressure. It is usually assumed that this does not matter. Does it matter in real life either? The question may be evaded by imagining microscopic plasticity smoothing away the microscopic roughness: but whether it does or not, it can be argued that the “contour area” behaves as if it were the contact area, and that the load and pressure distribution over the contact area are determined by the geometry of the contour area, not by the detailed variation within it. This returns the theory back to Archard’s idea (Whitehouse and Archard 1970) that only the “main scale roughness” need be measured, and reverses the traditional picture that initial elastic deformation is succeeded by plastic deformation (as in a Hertzian contact), instead, initial (small-scale) plastic deformation is succeeded by (large-scale) elastic deformation.

Persson’s Diffusion Theory of Elastic Contact

A completely different theory of rough surface contact has been proposed by Persson (2001), still based on the postulate of a Gaussian surface, but now focusing on the behavior as the resolution is improved, i.e., as the sampling interval is reduced, so that smaller and smaller surface features are seen. While traditional contact mechanics of smooth surfaces studies *spatial* pressure distributions, Persson studies the *probability distribution* of pressures – and finds that the distribution spreads as the sampling interval is reduced just as the temperature diffuses in a solid. As described above, the second moment of

the height distribution $m_2 = \pi \int_0^\infty k^3 \Phi^s(k) dk$ is often infinite; one way of accounting for finite resolution is to take the lower limit as some non-zero value k_1 related to the sampling interval. Now in elastic contact the pressure needed to flatten a sinusoidal height waviness $z = a(k) \sin(kx)$ is $p = \frac{1}{2} k E^* a(k) \sin(kx)$; by superposition it follows that the mean square of the pressures needed to flatten any rough surface (assembled from uncorrelated waves) is just $(\frac{1}{2} E^*)^2$ multiplying the mean square slope m_2 . Then as k_1 is reduced, the second moment increases without limit, and so does the mean square pressure $V \equiv \pi (\frac{1}{2} E^*)^2 \int_{k_1}^\infty k^3 \Phi^s(k) dk$, which can be regarded as a time variable in the diffusion equation $\frac{\partial p}{\partial V} = \frac{1}{2} \frac{\partial^2 p}{\partial p^2}$ giving the probability distribution P of the pressure p just as temperature distributions are governed by $\frac{\partial \theta}{\partial t} = \kappa \frac{\partial^2 \theta}{\partial x^2}$. Calculated pressure distributions for surfaces with different cut-offs show reasonable agreement with predictions, but are not completely convincing because of the difficulty of generating numerically a surface clearly satisfying the theoretical requirements, and the fact that the finite element calculations necessarily apply point loads at nodes, making the calculation of contact area and thence pressure slightly suspect. The need for verification is real, for the derivation of Persson’s equation has been challenged (Manners and Greenwood 2006) on the grounds that at one step complete contact is assumed (as in the elastic argument just above), but the equation is then applied to incomplete contact. Certainly this appears to be a valuable approach to the problem of contact of fractal surfaces, facing up to the inherent resolution dependence.

Cross-References

- [Fractal Characterization of Surfaces](#)
- [Fractal Contact Mechanics](#)
- [Fractal Geometry](#)
- [Fractal Nature of Surfaces](#)
- [Stochastic Models for Rough Surface EHL](#)

References

- A.W. Bush, R.D. Gibson, T.R. Thomas, The elastic contact of a rough surface. *Wear* **35**, 87–111 (1975)
- G. Carbone, F. Bottiglione, Asperity contact theories: do they predict linearity between contact area and load? *J. Mech. Phys. Solids* **56**, 2555–2572 (2008)
- J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces. *Proc. R. Soc. A* **295**, 300–319 (1966)
- J.A. Greenwood, A unified theory of surface roughness. *Proc. R. Soc. A* **393**, 133–157 (1984)

- J.A. Greenwood, A simplified elliptic model of rough surface contact. *Wear* **261**, 191–200 (2006)
- M.S. Longuet-Higgins, The statistical analysis of a random moving surface. *Phil. Trans. R. Soc. Lond.* **A249**, 321–337 (1957a)
- M.S. Longuet-Higgins, Statistical properties of an isotropic random surface. *Phil. Trans. R. Soc. Lond.* **A250**, 157–174 (1957b)
- W. Manners, J.A. Greenwood, Some observations on Persson's diffusion theory of elastic contact. *Wear* **261**, 600–610 (2006)
- P.R. Nayak, Random process model of rough surfaces. *J. Lub. Technol. (ASME)* **93**, 398–407 (1971)
- P.R. Nayak, Some aspects of surface roughness measurement. *Wear* **26**, 165–174 (1973)
- B. Persson, Theory of rubber friction and contact mechanics. *J. Chem. Phys.* **115**, 3840–3861 (2001)
- R.S. Sayles, T.R. Thomas, Surface topography as a non-stationary random process. *Nature* **271**, 431–434 (1978)
- R.S. Sayles, T.R. Thomas, Measurements of the statistical microgeometry of engineering surfaces. *J. Lub. Technol. (ASME)* **101**, 409–417 (1979)
- T.R. Thomas, *Rough Surfaces*, 2nd edn. (Imperial College Press, London, 1999)
- D.J. Whitehouse, J.F. Archard, The properties of random surfaces of significance in their contact. *Proc. R. Soc.* **A316**, 97–121 (1970)
- D.J. Whitehouse, M.J. Phillips, Two-dimensional discrete properties of random surfaces. *Phil. Trans. R. Soc. Lond.* **A305**, 441–468 (1982)

Stochastic Modeling of Flows in Lubrication

► Average Reynolds Equation

Stochastic Models for Rough Surface EHL

DONG ZHU

State Key Laboratory of Mechanical Transmission,
Chongqing University, Chongqing, People's Republic
of China

Synonyms

EHL modeling considering surface roughness; Rough surface EHL simulation with statistic models

Definition

Elastohydrodynamic lubrication (EHL) is a mode of fluid-film lubrication in which many mechanical components operate. Since roughness of engineering surfaces is usually of the same order of magnitude as, or greater than, EHL film thickness, its influence ought to be taken into account while investigating EHL characteristics. The roughness

effect on the EHL can be simulated with stochastic or deterministic models. This essay describes the stochastic models of rough surface EHL.

Scientific Fundamentals

Introduction

EHL (elastohydrodynamic lubrication) is a mode of fluid-film lubrication in which hydrodynamic action is significantly enhanced by surface elastic deformation and lubricant viscosity increase due to high pressure. Many mechanical components operate in the EHL regime. These include various gears, rolling element bearings, cam/follower systems, vane pumps, ball screws, metal-rolling tools, traction drives and continuously variable transmissions, and so on. A good understanding of EHL characteristics is vital to improvements of components performance, efficiency, and durability, as well as design optimization and failure prevention. Over the last 30–60 years, fundamental EHL theories have been established, numerical solution methods developed, and various film thickness formulae derived through curve-fitting based on obtained numerical solutions. Refer to “► [Elastohydrodynamic Lubrication \(EHL\)](#),” “► [EHL Film Thickness Behavior](#),” “► [Film Thickness Formulas: Line Contacts](#)” and “► [Film Thickness Formulas: Point Contacts](#)” for details. Also refer to Dowson and Higginson (1966) and Hamrock and Dowson (1977).

However, conventional EHL theories and various film thickness formulae were developed based on an assumption that both contacting surfaces are ideally smooth. In reality, roughness of engineering surfaces is usually of the same order of magnitude as, or greater than, the EHL film thickness, so the roughness effect ought to be taken into account while investigating the EHL characteristics.

Great efforts have been made by engineers and researchers to develop rough surface EHL models since 1970s. Basically, there have been two types of analysis taking into account the surface roughness effect: stochastic models and deterministic models. Early studies were focused on the stochastic models. L.S. H. Chow and H.S. Cheng (1976), developed a simplified EHL solution of Grubin's type, using a stochastic model for two-dimensional purely transverse or longitudinal roughness. H.S. Cheng and A. Dyson (1978), completed a numerical solution for the inlet half of an EHL line contact, employing a stochastic Reynolds equation for ideal longitudinal roughness and a load-compliance relation derived based on the statistic data obtained from a pair of circumferentially ground disks. Shortly afterwards, an average Reynolds equation was developed by N. Patir

and H.S. Cheng (1978a), and pressure and shear flow factors were determined through numerical flow simulations. Based on the average flow model and a simplified stochastic contact model by J.A. Greenwood and J.H. Tripp (1971), the effect of surface roughness and its orientation on the line contact EHL was investigated by Patir and Cheng (1978b), through a simplified solution for the inlet half of the EHL line contact. It was observed that EHL central film thickness could be increased with transversely oriented roughness, but reduced with longitudinal roughness. This effect could be significant when the hydrodynamic roughness parameter, $\Lambda = h_{cs}/\sigma$, is small, or insignificant if Λ is greater than 2.5~3. Patir and Cheng's model has been found to be helpful for basic understanding of rough surface lubrication. It also provided a simple mathematical tool for analyzing roughness effect on lubrication.

On the basis of the average flow model and stochastic contact model mentioned above, full numerical EHL solutions have been developed. Line contact solutions were presented by B.C. Majumdar and B.J. Hamrock (1982), and J. Prakash and H. Czichos (1983), and later by D. Zhu et al. (1990). A point contact solution was published by Zhu and Cheng (1988). Most line contact solutions have demonstrated the same characteristics as those by Patir and Cheng. For point contact with a large contact ellipticity greater than 2~3 (which is close to the line contact), the same trend has also been observed.

It is important to note that the stochastic models simulate the global effect of surface roughness and topography, and predict average values of hydrodynamic and asperity contact pressures and film thickness. In stochastic analyses detailed information about parameter distributions and local peaks, which may often be critical for the study of lubrication breakdown and surface failures, are missing. In order to obtain detailed parameter fluctuations, deterministic EHL models based on digitized real machined surfaces have been developed since the mid-1990s (refer to "► [Deterministic Models of Rough Surface EHL](#)"). However, deterministic analyses are usually complicated and time consuming, and the stochastic models are simple and more efficient.

This essay describes mainly the models and sample results presented by Zhu and Cheng (1988), for point contacts, as the point contact solution is more generic, showing fundamentals that are also representative of the line contact. Other line contact solutions mentioned above employed the same average Reynolds equation and simplified stochastic contact model by Greenwood and Tripp, so they will not be described here.

Basic Models

For a full numerical solution of the rough surface EHL problem in point contacts, the mean variables to be solved are the mean hydrodynamic pressure, $p(\mathbf{x}, \mathbf{y})$, the mean asperity contact pressure, $p_a(\mathbf{x}, \mathbf{y})$, and the nominal film thickness, $h(\mathbf{x}, \mathbf{y})$. The governing equation for the mean hydrodynamic pressure under steady-state condition, as presented by Patir and Cheng (1978a), is written as follows:

$$\frac{\partial}{\partial x} \left(\Phi_x \frac{\rho h^3}{12\eta} \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial y} \left(\Phi_y \frac{\rho h^3}{12\eta} \frac{\partial p}{\partial y} \right) = \frac{u_1 + u_2}{2} \frac{\partial(\rho \bar{h}_T)}{\partial x} + \frac{u_1 - u_2}{2} \sigma \frac{\partial(\rho \Phi_s)}{\partial x}$$

where Φ_x and Φ_y are pressure flow factors, Φ_s shear flow factor, σ standard deviation of surface roughness, and \bar{h}_T average gap. These flow factors can be obtained either through numerical flow simulation, or calculated with empirical formulae by Patir and Cheng (1978a), which will not be repeated here. The average gap is calculated by

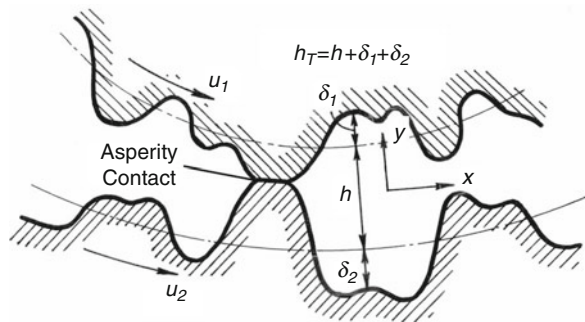
$$\bar{h}_T = \int_{-h}^{\infty} (h + \delta) f(\delta) d\delta$$

where $f(\delta)$ is the probability density function of combined roughness δ . For a Gaussian distribution, which can often be found from machined engineering surfaces, it can be expressed as (Fig. 1)

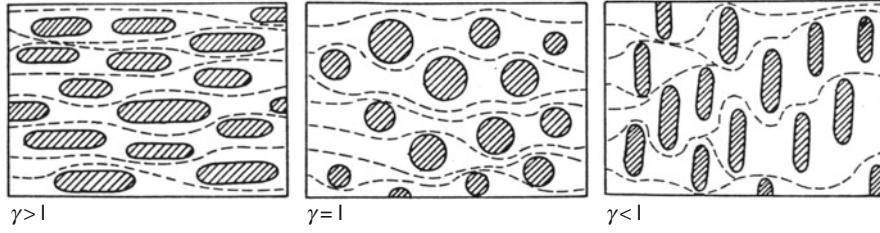
$$f(\delta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\delta^2}{2\sigma^2}}$$

When the average film thickness is small, the roughness effect is significant and the asperity contacts may occur. Figure 2 shows typical flow and asperity contact patterns from different surface orientations. The surface pattern parameter, γ , is defined as:

$$\gamma = \lambda_{0.5x} / \lambda_{0.5y}$$



Stochastic Models for Rough Surface EHL, Fig. 1 Film thickness/gap function



Stochastic Models for Rough Surface EHL, Fig. 2 Typical hydrodynamic flow and asperity contact patterns for longitudinal ($\gamma > 1$), isotropic ($\gamma = 1$), and transverse ($\gamma < 1$) roughness

where $\lambda_{0.5x}$ and $\lambda_{0.5y}$ are correlation lengths in x - and y -directions, at which the auto-correlation function of the profile reduces to 50% of its initial value. The influence of surface pattern is considered in the calculations of flow factors, Φ_x and Φ_y (see Patir and Cheng 1978a, for details).

In the case of general point contact, the contact zone is an ellipse and its geometry can be determined according to the Hertzian theory. The nominal EHL film thickness can be given below:

$$h(x, y) = h_c + \frac{x^2}{2R_x} + \frac{y^2}{2R_y} + V(x, y) - V(0, 0)$$

where h_c is the nominal central film thickness and R_x and R_y are effective radii of curvature in the x - and y -directions, respectively. The bulk surface deformation at point (x, y) is governed by total normal pressure, which is the sum of the mean hydrodynamic and asperity contact pressures, so it can be computed by

$$V(x, y) = \frac{2}{\pi E'} \iint_{\Omega} \frac{p(\xi, \zeta) + p_c(\xi, \zeta)}{\sqrt{(x - \xi)^2 + (y - \zeta)^2}} d\xi d\zeta$$

For a piezo-viscous lubricant, various pressure-viscosity equations have been developed. The following Barus viscosity model has been commonly used (refer to “► EHL Governing Equations” for details):

$$\eta = \eta_o e^{\alpha p}$$

Perhaps the most uncertain quantity in the stochastic EHL analysis is the asperity contact pressure. Although the force-compliance relation has been studied by many researchers, a general theory applicable to different rough surfaces is not available. In the region of $\Lambda = h/\sigma > 0.5$, however, the asperity contact pressure is usually a small portion of the total pressure, so that the use of an approximate relationship is justified. One of the simplified relations based on stochastic contact analyses is given below, which was developed by J.A. Greenwood and J.H. Tripp (1971):

$$p_a = K' E' F_{2.5}(\lambda)$$

where $K' = \frac{8}{15} \sqrt{2\pi} (N\beta\sigma) \sqrt{\sigma/\beta}$, N is the number of asperities per unit area, β the mean radius of curvature of asperities, and $F_{2.5}$ is a function of λ defined as follows:

$$F_{2.5} = \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} (S - \lambda)^{2.5} e^{-\frac{S^2}{2}} dS$$

After obtaining both hydrodynamic pressure and asperity contact pressure distributions in the solution domain, the total load can be calculated by

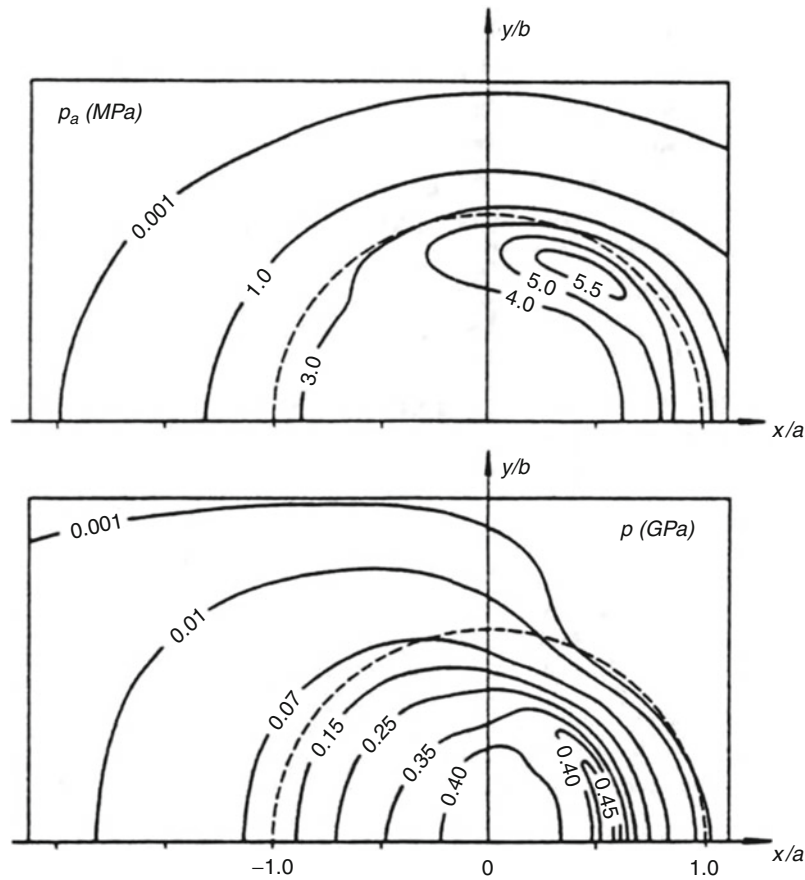
$$W = W_h + W_a = \iint_{\Omega} p(x, y) dx dy + \iint_{\Omega} p_a(x, y) dx dy$$

where W_h and W_a are hydrodynamic load and contact load, respectively. The contact load ratio, W_o is defined as $W_o = W_a/W$, which describes load sharing and contact severity in mixed lubrication.

Sample Results and Discussion

A direct iterative procedure was employed to solve the equation system given above, and sample cases were analyzed at $N\beta\sigma = 0.04$ on both surfaces (which is typical in engineering practice) under pure rolling conditions, so the shear flow term in the Reynolds equation vanishes. In order to investigate the effects of roughness and its orientation, roughness pattern parameter γ was changed from 1/9, 1/6, 1/3, 1, 3, ... up to 9, and composite RMS roughness σ adjusted to have the hydrodynamic roughness parameter $\Lambda = h_{cs}/\sigma$ varying in a wide range from 0.5 to 6.0. Figure 3 illustrates a typical solution at $\Lambda = 0.583$, in which considerable asperity contacts were observed and the contact pressure distribution $p_a(x, y)$ correlates well with that of the nominal film thickness $h(x, y)$. However, compared with the hydrodynamic pressure, asperity contact pressure is still quite limited even at the low roughness parameter of $\Lambda = 0.583$.

In Fig. 4, the ratio of the nominal central film thickness of rough surfaces to that of smooth surfaces, h_c/h_{cs} , is plotted for various γ and Λ , in comparison with results from Patir and Cheng (1978b), and Zhu et al. (1990).



Stochastic Models for Rough Surface EHL, Fig. 3 Mean asperity contact and hydrodynamic pressure distributions in a mixed EHL elliptical contact $U^* = 0.8414 \times 10^{-11}$, $G^* = 2,637$, $W^* = 0.9211 \times 10^{-6}$

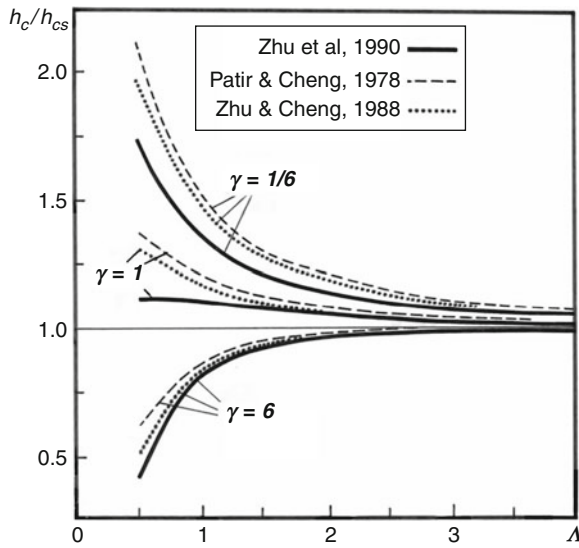
Although three different studies were conducted under different operating conditions, the results demonstrate the same trends, i.e., at low Λ values the roughness effect appears to be significant, but when $\Lambda > 2 \sim 3$ the effect becomes very limited or negligible. Also, transverse roughness, $\gamma < 1.0$, yields a higher film thickness than those from $\gamma \geq 1.0$, the isotropic and longitudinal roughness. Note that the results from Zhu et al. (1990), were from line contact cases under heavy loading condition, and those from Zhu and Cheng (1988), were from cases in a elliptical contact of $K = 3.955$, which is close to line contact.

Figure 5 shows the effects of surface roughness parameters, γ and Λ , on the load sharing between the hydrodynamic film and the asperity contacts. It can be seen that the contact load ratio $W_c = W_a/W$ increases as Λ decreases or γ increases. But even in the case of Λ as small as 0.5 and γ as large as 9.0, the contact load is still a small portion of the total load. Similar results were observed in other

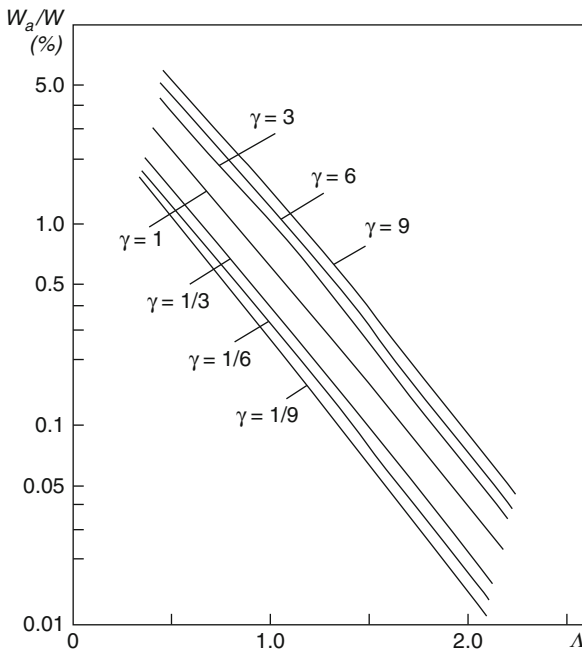
studies, such as those by Patir and Cheng (1978b), and Majumdar and Hamrock (1982), and others. This is due to intrinsic limitations from the stochastic contact model used, which does not allow deep penetration/interaction of asperities and also may underestimate the contact pressure in many cases. Recent studies after the mid-1990s using deterministic models with digitized machined roughness have yielded results qualitatively with the same trends, but cases quantitatively with much severe asperity contacts have been analyzed. Refer to “► [Deterministic Models of Rough Surface EHL](#)” for details.

Key Applications

EHL is an important branch of lubrication theory, describing lubrication mechanisms in non-conformal contacts, which can be widely found in many mechanical components such as various gears, rolling bearings, cams and followers, hydraulic vane pumps, ball screws, traction drives



Stochastic Models for Rough Surface EHL, Fig. 4 Effects of surface pattern parameter γ and hydrodynamic roughness parameter Λ on nominal central film thickness



Stochastic Models for Rough Surface EHL, Fig. 5 Load sharing as a function of roughness parameters, γ and Λ

and continuous variable transmissions, and metal rolling tools. These components usually transmit substantial power and motion, and they are often heavily loaded. Since engineering surfaces made by common machining

processes are often quite rough, and the surface roughness is usually of the same order of magnitude as, or greater than, the EHL film thickness, the surface roughness effect ought to be considered while investigating the contact and EHL characteristics. There have been mainly two types of models analyzing the surface roughness effect: stochastic and deterministic models. Stochastic approaches are relatively simple and efficient, but they only provide mean values of lubricant film thickness and hydrodynamic and contact pressures, demonstrating global trends of lubrication performance and transition. Deterministic models are able to provide more detailed information about the film thickness and pressure fluctuations, but they are usually complicated and time-consuming. Stochastic models, therefore, are practically useful for predicting roughness effect and lubrication performance. They can provide valuable information in engineering practice for improving component functionality, efficiency and durability, especially for conformal contact components, such as journal bearings and pistons, with large areas of surface interaction that may so far still be difficult to handle by deterministic models.

Nomenclature

a	Semi-axis of Hertzian ellipse in x-direction, or radius of Hertzian circle
b	Semi-axis of Hertzian contact ellipse in y-direction
E'	Effective elastic modulus
G^*	$\alpha E'$ dimensionless material parameter
h	Nominal local film thickness (or gap)
h_c	Nominal central film thickness
h_{cs}	Central film thickness predicted by smooth surface EHL theory
h_m	Nominal minimum film thickness
K	b/a Hertzian contact ellipticity
p	Mean hydrodynamic pressure
p_a	Mean asperity contact pressure
p_h	Maximum Hertzian pressure
R_q	Root mean square (RMS) surface roughness
R_x, R_y	Effective radii of curvature in x- and y-directions
S	$(u_2 - u_1)/U$ slide-to-roll ratio
U^*	$\eta_0 U / (E' R_x)$ dimensionless speed parameter
U	$(u_1 + u_2)/2$ rolling velocity (or entraining velocity)
u_1, u_2	Velocities of Surface 1 and Surface 2, respectively
W^*	$W / (E' R_x^2)$ for point contact, or $w / (E' R_x l_e)$ for line contact dimensionless load parameter
W	Total load
W_a	Contact load

W_h	Hydrodynamic load
W_c	W_d/w contact load ratio
x, y	Coordinates (x is chosen to be parallel to rolling direction)
α	Pressure–viscosity exponent used in pressure–viscosity equation $\eta = \eta_o EXP(\alpha p)$
η, η_o	Viscosity and viscosity under ambient condition, respectively
A	h_{cs}/σ hydrodynamic roughness parameter
λ	h/σ
s	$(R_{q1}^2 + R_{q2}^2)^{0.5}$ composite RMS roughness

Cross-References

- [Average Reynolds Equation](#)
- [Deterministic Models of Rough Surface EHL](#)
- [Elastohydrodynamic Lubrication \(EHL\)](#)
- [Film Thickness Formulas: Line Contacts](#)
- [Film Thickness Formulas: Point Contacts](#)
- [Mixed EHL](#)
- [Point Contact EHL](#)

References

- H.S. Cheng, A. Dyson, Elastohydrodynamic lubrication of circumferentially-ground rough disks. *ASLE Trans.* **21**, 25–40 (1978)
- L.S.H. Chow, H.S. Cheng, The effect of surface roughness on the average film thickness between lubricated rollers. *J. Lubr. Technol.* **98**, 117–124 (1976)
- D. Dowson, G.R. Higginson, *Elastohydrodynamic Lubrication* (Pergamon, London, 1966)
- J.A. Greenwood, J.H. Tripp, The contact of two nominally flat rough surfaces. *Proc. Inst. Mech. Eng.* **185**(48 Pt. 1), 625–633 (1970/1971)
- B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 3 – fully flooded results. *J. Lubr. Technol.* **99**, 264–276 (1977)
- B.C. Majumdar, B.J. Hamrock, Effect of surface roughness on elastohydrodynamic line contact. *J. Lubr. Technol.* **104**, 401–409 (1982)
- N. Patir, H.S. Cheng, An average flow model for determining effects of three-dimensional roughness on partial hydrodynamic lubrication. *J. Lubr. Technol.* **100**, 12–17 (1978a)
- N. Patir, H.S. Cheng, Effect of surface roughness orientation on the central film thickness in EHD contacts, in *Proceedings of the 5th Leeds-Lyon Symposium on Tribology*, 1978b, pp. 15–21
- J. Prakash, H. Czichos, Influence of surface roughness and its orientation on partial elastohydrodynamic lubrication of rollers. *J. Lubr. Technol.* **105**, 591–597 (1983)
- D. Zhu, H.S. Cheng, Effect of surface roughness on the point contact EHL. *ASME J. Tribol.* **110**, 32–37 (1988)
- D. Zhu, H.S. Cheng, B.J. Hamrock, Effect of surface roughness on pressure spike and film constriction in elastohydrodynamically lubricated line contacts. *Tribol. Trans.* **33**, 267–273 (1990)

Stokes Equation and Its Application in Lubrication

D. C. SUN

Department of Mechanical Engineering, State University of New York at Binghamton, Binghamton, NY, USA

Definition

The Stokes equation is a special case of the Navier-Stokes equation and describes the very slow motion of a viscous fluid. The Stokes equation may be applied to analyze the surface roughness effect in fluid film lubrication when the roughness spacing is not much greater than the lubricant film thickness.

Scientific Fundamentals

Stokes Equation

In the very slow motion of a viscous fluid the inertia of the fluid may be neglected in comparison with the other forces acting on the fluid. Since the Reynolds number represents the ratio of the inertia forces to the viscous forces, such a flow is called the low-Reynolds-number flow. The Stokes equation is obtained by neglecting all the convective inertia terms in the Navier-Stokes equation (Currie 1974). In usual applications the density is constant, the body force is absent, and the Stokes equation takes the form:

$$\rho \frac{\partial \mathbf{V}}{\partial t} = -\nabla p + \eta \nabla^2 \mathbf{V} \quad (1)$$

where \mathbf{V} is the velocity of flow, p is the pressure, ρ is the density, η is the dynamic viscosity, and t is the time.

Fluid Film Lubrication

The contact between two solid surfaces in relative motion can be lubricated by introducing a thin fluid film between them. The film thickness, h , is typically about $L/1,000$, where L characterizes the size of the contact region. The thin-film flow can be described by a set of equations obtained from further simplifying (1):

$$0 = -\frac{\partial p}{\partial x} + \eta \frac{\partial^2 u}{\partial z^2} \quad (2a)$$

$$0 = -\frac{\partial p}{\partial y} + \eta \frac{\partial^2 v}{\partial z^2} \quad (2b)$$

$$0 = -\frac{\partial p}{\partial z} \quad (2c)$$

where (x, y) are the spatial coordinates in the plane of the thin film, z is the spatial coordinate perpendicular to the plane of the thin film, and (u, v) are the x - and y -components of \mathbf{V} . Osborne Reynolds first envisaged this simplification and derived from it an equation, which has since been known as the Reynolds equation, that explained the generation of pressure in the film. Reynolds theory of fluid film lubrication reveals a crucially important lubrication mechanism and is a marvelous triumph of fluid mechanics.

Surface Roughness Effect

The description of surface roughness encompasses two categories of properties (ANSI 1978), one associated with the roughness height, the other associated with the roughness spacing. The surface roughness effect in fluid film lubrication needs to be considered when the roughness height is significant relative to the film thickness (say, larger than 30% of it). Naturally, the Reynolds equation is used to analyze the roughness effect. Due to the inherent simplification involved in its derivation, however, the use of the Reynolds equation is appropriate only if the roughness spacing is much larger (say, more than 10 times) than the film thickness. If the roughness spacing is not much larger than the film thickness (say, the two being of similar magnitudes), then certain neglected terms (hereafter to be referred to as the neglected terms) from (1) to arrive at (2a–2c) become significant, and (1) needs to be used instead to analyze the roughness effect. Elrod (1973) proposed to name two kinds of roughness, “Reynolds roughness” or “Stokes roughness,” depending upon whether the Reynolds equation or the Stokes equation, respectively, applies for analyzing their effect. Thus, the Stokes equation finds an application in fluid film lubrication.

Stochastic Description of Surface Roughness

The detailed geometry of a rough surface is usually so irregular that it requires a statistical representation. Let the film thickness be written as:

$$h(x, y) = h_d(x, y) + h_r(x, y) \quad (3)$$

where h_d is the nominal film thickness (the deterministic component) and h_r is the roughness height (the random component). The random component may then be modeled as an ergodic, weakly stationary random process of zero mean (Papoulis 1965). Among the numerous statistical variables, it is adequate and practical to consider only the autocorrelation function (ACF):

$$R(|x - x_1|, |y - y_1|) = \langle h_r(x, y) h_r(x_1, y_1) \rangle \quad (4)$$

where (x, y) and (x_1, y_1) are two separate points on the rough surface, and the bracket $\langle \rangle$ means ensemble average.

Through the ACF the randomness and periodicity of a roughness profile can be represented by the correlation length (β) and the correlation wavelength (γ), respectively. Both these lengths are measures of the roughness spacing. By merging the two points in the ACF one also obtains the variance,

$$R(0, 0) = \langle h_r^2 \rangle \equiv \sigma^2 \quad (5)$$

which is the rms-roughness-height (σ) squared. Thus, the ACF describes both the roughness height and the roughness spacing, and it has been recognized as an effective tool for surface characterization (Peklenik 1967).

Effect of Stokes Roughness

The Stokes roughness effect was analyzed with an iterative method (Sun and Chen 1977). The method is to solve (2a–2c) and (3) first to obtain a solution containing the (Reynolds) roughness effect. The neglected terms from (1) are evaluated with this solution to become additional terms in (2a–2c), which then are solved again. Essentially the process is the substitution of the Reynolds roughness solution into the neglected terms and the evaluation of their influence. The iterative solution is further expanded by assuming the roughness height to be small relative to the nominal film thickness, and only terms up to the second order are retained. The roughness height h_r is considered to vary only in the direction of relative sliding between the two surfaces (the transverse roughness [the type of roughness where the undulations of the surface occur only in the sliding direction is called the “transverse roughness,” whereas, the one with the undulations occurring only in the direction perpendicular to sliding is called the “longitudinal roughness” (Christensen 1969–70)]).

The model of lubrication is a slider bearing with an exponential film thickness function:

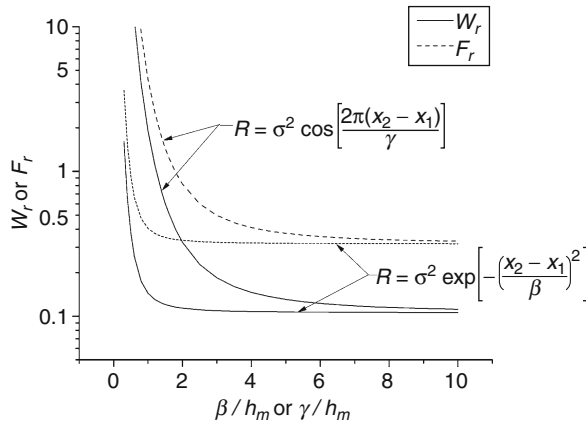
$$h_d(x) = h_m \exp \left[1 - \frac{x}{L} \right], 0 \leq x \leq L \quad (6)$$

where x is in the direction of sliding (with speed U), L is the slider length, and h_m is the minimum film thickness. Two types of ACF are considered: If $h_r(x)$ is a wide band random noise, the ACF is modeled as:

$$R(|x - x_1|) = \sigma^2 \exp \left[- \left(\frac{x - x_1}{\beta} \right)^2 \right] \quad (7a)$$

If $h_r(x)$ is a sine wave without any random distortion, the ACF is modeled as:

$$R(|x - x_1|) = \sigma^2 \cos \left[\frac{2\pi(x - x_1)}{\gamma} \right] \quad (7b)$$



Stokes Equation and Its Application in Lubrication, Fig. 1
Stokes roughness effect

The result of the analysis may be summarized as follows (Sun and Chen 1977):

$$W = W_d + \left(\frac{\sigma}{h_m}\right)^2 W_r \quad (8)$$

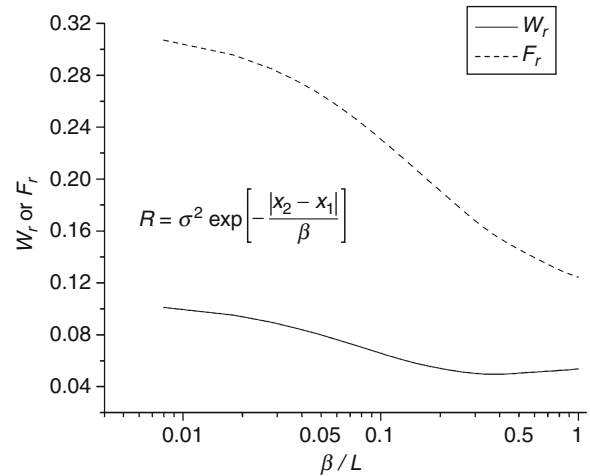
$$F = F_d + \left(\frac{\sigma}{h_m}\right)^2 F_r \quad (9)$$

where W is the load-capacity per unit width of the slider, normalized by $6\eta UL^2/h_m^2$, and F is the sliding friction per unit width of the slider, normalized by $6\eta UL/h_m$. For the film thickness function given in (6), the smooth-surface load-capacity is $W_d = 0.027$ and the smooth-surface sliding friction is $F_d = 0.126$. The second terms in (8) and (9) represent the roughness effect, whose magnitude is seen to be proportional to $(\sigma/h_m)^2$. The factors W_r and F_r depend on the shape of the ACF and are shown in Fig. 1. From the figure it is seen that (1) roughness causes both the load-capacity and friction to increase above their smooth-surface values, (2) the effect becomes greater as the correlation length or correlation wavelength decreases toward the film thickness, and (3) roughness of the sinusoidal wave type produces more pronounced effect than the wide band random noise type.

Comparison with Reynolds Roughness Effect

The effect of two-dimensional Reynolds roughness was analyzed (Sun 1978) for an ACF of the wide band random noise type:

$$R(|x - x_1|, |y - y_1|) = \sigma^2 \exp \left[-\frac{|x - x_1|}{\beta_x} - \frac{|y - y_1|}{\beta_y} \right] \quad (10)$$



Stokes Equation and Its Application in Lubrication, Fig. 2
Reynolds roughness effect

where β_x and β_y are the correlation lengths in the x - and y -directions, respectively. The model of lubrication is the same slider bearing with its film thickness function given in (6). A complete set of results showing the effects of the transverse roughness, the longitudinal roughness, and any combination of (β_x, β_y) values, can be found in (Sun 1978). For the sake of comparison with the Stokes roughness effect, the result for the special case of the transverse roughness is shown in Fig. 2. By comparing Fig. 1 with Fig. 2, it is seen that (1) the Stokes roughness effect is generally much greater than the Reynolds roughness effect, and (2) as the correlation length becomes large relative to the film thickness (from the Stokes roughness side) and becomes small relative to the bearing length (from the Reynolds roughness side), the two sets of curves mesh into each other.

Cross-References

- Average Reynolds Equations
- Fractal Characterization of Surfaces
- Fractal Nature of Surfaces
- Homogenization of the Reynolds Equation
- Navier-Stokes Equation and Applications in Lubrication
- Reynolds Equation
- Stochastic Models for Rough Surface EHL
- Surface Roughness
- Surface Texture Generation with a Numerical Process

References

- H. Christensen, Stochastic Models for Hydrodynamic Lubrication of Rough Surfaces. *Proc. Inst. Mech. Eng.* **181**(55), 1013–1026 (1969–70). part 1
- I.G. Currie, *Fundamental Mechanics of Fluids* (McGraw-Hill, New York, 1974), 253–255
- H.G. Elrod, Thin-film lubrication theory for Newtonian fluids with surfaces possessing striated roughness or grooving. *J. Lubr. Technol.* **95**, 484–489 (1973)
- A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1965). Ch. 9
- J. Peklenik, Investigation of the surface typology. *Ann. CIRP* **15**, 381–385 (1967)
- D.C. Sun, K.K. Chen, First effects of stokes roughness on hydrodynamic lubrication. *J. Lubr. Technol.* **99**, 2–9 (1977)
- D.C. Sun, On the effects of two-dimensional reynolds roughness in hydrodynamic lubrication. *Proc. R. Soc. Lon. A.* **364**, 89–106 (1978)
- The American National Standards Institute, *Surface Texture (ANSI B46.1)*, (ASME, New York, 1978)

Straight Gears

► [Spur Gears](#)

Straight-Cut Gears

► [Spur Gears](#)

Strain-Life Theories

WEICHENG CUI, FANG WANG
China Ship Scientific Research Center,
Wuxi Jiangsu, People's Republic of China

Synonyms

[Coffin-Manson law](#); ϵ - N curve

Definition

In principle, strain life is the same as stress life if the stress-strain relation is monotonic. However, when the stress exceeds the proportional limit, the stress is not unique for describing the material state, and the strain is better for describing the material state in this condition. Thus, strain-life theory has been developed. The strain-based approach to fatigue considers the plastic deformation

that may occur in localized regions where fatigue cracks begin, as at edges of beams and at stress raisers. Stresses and strains in such regions are analyzed and used as a basis for life estimates. This procedure permits detailed consideration of fatigue situations where local yielding is involved, which is a comprehensive approach that can be used both at short and long lives but more often, in the case for ductile metals, at relatively short lives. This approach was initially developed in the late 1950s and early 1960s in response to the need to analyze the fatigue problems involving fairly short fatigue lives.

Scientific Fundamentals

ϵ - N Curves

A strain versus life curve (ϵ - N curve) is a plot of strain amplitude versus cycles to failure. Such a curve is employed in the strain-based approach for making life estimates.

In most practical cases of fatigue design, the critical location will be a notch in which plastic strains are imposed by the surrounding elastic material. Thus, the situation will be strain controlled with a total strain range composed of elastic and plastic parts as

$$\frac{\Delta \epsilon_T}{2} = \frac{\Delta \epsilon_e}{2} + \frac{\Delta \epsilon_p}{2} \quad (1)$$

where $\frac{\Delta \epsilon_T}{2}$, $\frac{\Delta \epsilon_e}{2}$, $\frac{\Delta \epsilon_p}{2}$ are the total strain amplitude, the elastic strain amplitude, and the plastic strain amplitude, respectively.

The elastic strain amplitude is related to stress amplitude by

$$\frac{\Delta \epsilon_e}{2} = \frac{\Delta \sigma}{2E} = \frac{\sigma'_f}{E} (2N)^b \quad (2)$$

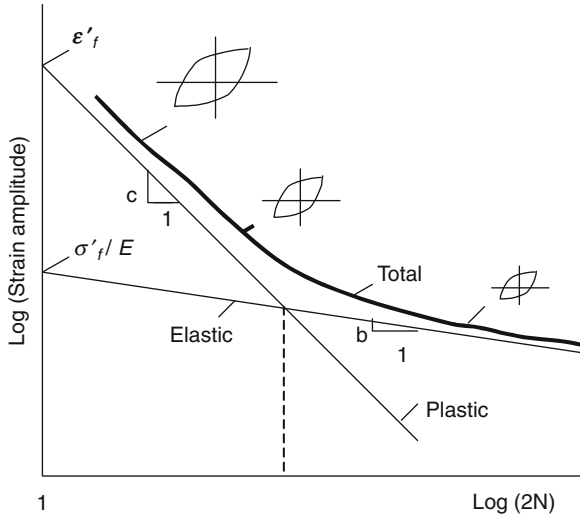
where $\frac{\Delta \sigma}{2}$ is the stress amplitude; E is the Young's modulus; N is the cycles to failure; $2N$ is load reversals to failure; σ'_f is the fatigue strength coefficient; and b is the fatigue strength exponent – the sign of b is negative.

The plastic strain amplitude is a measure of the half-width of the stress-strain hysteresis loop. The plastic strain resistance is best described by the so-called Manson-Coffin relationship:

$$\frac{\Delta \epsilon_p}{2} = \epsilon'_f \cdot (2N)^c \quad (3)$$

where ϵ'_f is the fatigue ductility coefficient and c is the fatigue ductility exponent – the sign of c is negative.

Manson and Hirschberg (1964) proposed that a metal's resistance to total strain cycling can be



Strain-Life Theories, Fig. 1 Representation of elastic, plastic, and total strain resistance to fatigue loading

considered as a superposition of its elastic and plastic strain resistance. By combining (2) and (3),

$$\frac{\Delta \varepsilon_T}{2} = \varepsilon_a = \frac{\Delta \varepsilon_e}{2} + \frac{\Delta \varepsilon_p}{2} = \frac{\sigma'_f}{E} (2N)^b + \varepsilon'_f \cdot (2N)^c \quad (4)$$

The total strain life curve approaches the plastic strain life curve in the low cycle region and the stress life curve in the high cycle region, as shown in Fig. 1.

Factors Affecting Strain-Life Curves

Surface Finish Effect

It is pointed out in Dowling (2007) that surface finish is important in high-cycle fatigue, because most of the life at the low stresses involved is spent initiating a crack. However, if significant plastic strains are present, a small crack (or crack-like damage) starts relatively early in the life, even if the surface is smooth. Most of the life is thus spent in growing small cracks into the material at some depth, where the surface finish cannot have an effect. A reasonable method of modifying the strain-life curve to include the slight effect of surface finish is to change only the elastic slope b by adding a component related to fatigue limit and a surface effect factor, while leaving fatigue strength coefficient unchanged:

$$b_s = b + \frac{\log m_s}{\log(2N_e)} \quad (5)$$

where N_e is the number of cycles associated with the fatigue limit and m_s is a surface factor.

Component Size Effect

Size effects are an important consideration in applying a strain-based approach to large members (Dowling 2007). The study on steel shaft diameter effects suggested lowering the entire strain-life curve by a factor m_d related to shaft diameter d in millimeters:

$$m_d = (d/25.4 \text{ mm})^{-0.093} \quad (6)$$

Then the coefficients σ'_f and ε'_f can be replaced by $m_d \sigma'_f$ and $m_d \varepsilon'_f$, while the slope constants b and c are not altered.

Mean Stress Effect

The strain-life curves are usually derived from the tests under completely reversed cyclic loading. The strain-life curve should be modified if the mean stress is changed.

The approach suggested by J. Morrow is the most famous one for steels to consider mean stress effect (Dowling 2007), which can be expressed as an equation giving the equivalent completely reversed stress amplitude, which is expected to produce the same life as a given combination of amplitude $\Delta\sigma/2$ and mean stress σ_m . Based on the approach proposed by Morrow, a single equation for the family of strain-life curves can be obtained (Raske and Morrow 1969):

$$\frac{\Delta \varepsilon_T}{2} = \frac{\sigma'_f}{E} \left(1 - \frac{\sigma_m}{\sigma'_f}\right) (2N)^b + \varepsilon'_f \left(1 - \frac{\sigma_m}{\sigma'_f}\right)^{c/b} (2N)^c \quad (7)$$

In order to reduce the estimated effect of mean stress at relatively short lives, the Morrow equation was modified to

$$\frac{\Delta \varepsilon_T}{2} = \frac{\sigma'_f}{E} \left(1 - \frac{\sigma_m}{\sigma'_f}\right) (2N)^b + \varepsilon'_f (2N)^c \quad (8)$$

The Morrow approach works quite well, but only for steels. Subsequently, the SWT (Smith, Watson, and Topper) equation was proposed, which provides acceptable results for a wide range of materials:

$$\sigma_{\max} \cdot \frac{\Delta \varepsilon_T}{2} = \frac{(\sigma'_f)^2}{E} (2N)^{2b} + \sigma'_f \varepsilon'_f (2N)^{b+c} \quad (9)$$

where $\sigma_{\max} = \sigma_m + \Delta\sigma/2$.

Multi-Axial Strain Effect

Engineering components are often subjected to complex multi-axial loading in which not only the amplitude of loading changes with time but also the principal axes rotate. Two approaches are usually used for this problem – the effective strain approach and the critical plane approach.

In the effective strain approach, the fatigue life for multi-axial loading is postulated to depend on the value of the effective strain amplitude ($\Delta\sigma/2$):

$$\overline{(\Delta\epsilon_T/2)} = \frac{\overline{(\Delta\sigma/2)}}{E} + \overline{(\Delta\epsilon_p/2)} \quad (10)$$

where the effective stress amplitude ($\overline{(\Delta\sigma/2)}$) and effective plastic strain amplitude ($\overline{(\Delta\epsilon_p/2)}$) can be obtained by substituting amplitudes of the principal stresses and plastic strains to the following two equations:

$$\bar{\sigma} = \frac{1}{\sqrt{2}} \sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2} \quad (11)$$

$$\bar{\epsilon}_p = \frac{\sqrt{2}}{3} \sqrt{(\epsilon_{p1} - \epsilon_{p2})^2 + (\epsilon_{p2} - \epsilon_{p3})^2 + (\epsilon_{p1} - \epsilon_{p3})^2} \quad (12)$$

where $\sigma_1, \sigma_2, \sigma_3$ are the principal normal stresses and $\epsilon_{p1}, \epsilon_{p2}, \epsilon_{p3}$ are the plastic strains in the principal direction.

Where the loading is non-proportional to a significant degree, a critical plane approach is needed. In such an approach, stresses and strains acting on the most severely loaded plane are determined for various orientations (planes) in the material, and the stresses and strains acting on the most severely loaded plane are used to predict fatigue failure. Under multi-axial fatigue loadings, some of the most successful criteria are based on the critical plane approach. In general, these methods are based on the combined use of the shear stress amplitude acting on the plane experiencing the maximum shear stress amplitude (critical plane) and the maximum stress normal to this plane (Socie and Marquis 2000).

For example, Chu (1995) suggested a single multi-axial fatigue criterion that considers both the shear and normal stress cracking modes:

$$2\tau_{\max} \frac{\Delta\gamma}{2} + \sigma_{\max} \frac{\Delta\epsilon}{2} = f(N_f) \quad (13)$$

where $\tau_{\max}, \frac{\Delta\gamma}{2}, \sigma_{\max}, \frac{\Delta\epsilon}{2}$ are maximum shear stress, shear strain amplitude, maximum normal stress, and normal strain amplitude, respectively. The critical plane is simply the plane where the left-hand side of (13) is largest, and $f(N_f)$ can be obtained from uniaxial test data.

Crack Growth Effect

In the usual manner of obtaining strain-life curves, the N values correspond to failure or to substantial cracking in small axial test specimens. It is generally observed that life predictions made on this basis correspond to an engineering size crack that is easily visible with the naked eye. The existence of such a crack is often considered to constitute failure of the component. However, this rather loose

definition of failure is not always sufficient. Strain-life curves corresponding to specific small crack sizes are needed if this is to be done in a rigorous manner.

Simplifications

For general low cycle and high cycle fatigue, the Manson-Coffin relationship, (4), has a strong curve-fit ability but it needs to determine five material properties. Manson (1965) has simplified the equation even further with his universal slope method, where

$$\Delta\epsilon = 3.5 \frac{\sigma_u}{E} (N)^{-0.12} + \epsilon_f^{0.6} (N)^{-0.6} \quad (14)$$

σ_u, E , and ϵ_f are all obtained from a monotonic tensile test. He assumed the two exponents are fixed for all materials and only σ_u, E , and ϵ_f control the fatigue behavior.

Mitchell et al. (1977) proposed that the exponent b is also a function of σ_u and estimated ϵ_f' directly from the true fracture ductility ϵ_f . It is assumed that Manson's slope $c = -0.6$ is only valid for "ductile" metals, while $c = -0.5$ should be more appropriate for "strong" metals. Then the equation is

$$\begin{aligned} \frac{\Delta\epsilon}{2} &= \frac{\sigma_u + 345}{E} (2N)^{\frac{1}{6} \log \frac{0.5\sigma_u}{\sigma_u + 345}} + \epsilon_f (N)^c \\ c &= \begin{cases} -0.6 & \text{for "ductile" metals} \\ -0.5 & \text{for "strong" metals} \end{cases} \end{aligned} \quad (15)$$

Later, Manson's equation was further modified to the following expression (Muralidharan and Manson 1988):

$$\begin{aligned} \Delta\epsilon &= 0.0266 D^{0.115} [\sigma_u/E]^{-0.53} N_f^{-0.56} \\ &\quad + 1.17 [\sigma_u/E]^{0.832} N_f^{-0.09} \end{aligned} \quad (16)$$

where σ_u is the ultimate strength of metal, D is the ductility of the metal, E is the modulus of elasticity, and N_f is the fatigue life. Good correlation between the fatigue life predicted by this equation and the fatigue test data has been found (Muralidharan and Manson 1988).

Based on a detailed correlation study between monotonic tensile data and constant amplitude strain-controlled fatigue properties, the following simple strain-life formula is proposed by Roessle and Fatemi (2000):

$$\begin{aligned} \frac{\Delta\epsilon}{2} &= \frac{4.25(HB) + 225}{E} (2N_f)^{-0.09} \\ &\quad + \frac{0.32(HB)^2 - 487(HB) + 191000}{E} (2N_f)^{-0.56} \end{aligned} \quad (17)$$

This approximation uses only the Brinnell hardness HB and modulus of elasticity E as inputs for strain-life

approximation, both of which are either commonly available or easily measurable.

The inclusion of mean stress or mean strain effects in fatigue life prediction methods involving strain-life data is very complex. One method is to replace σ'_f with $\sigma'_f - \sigma_m$ in (14), where σ_m is the mean stress such that

$$\frac{\Delta \varepsilon}{2} = \frac{(\sigma'_f - \sigma_m)(2N)^b}{E} + \varepsilon'_f(2N)^c \quad (18)$$

where σ_m is taken positive for tensile values and negative for compressive values. Another equation suggested by Smith et al. (1970), based on strain-life test data at fracture obtained with various mean stresses, is

$$\sigma_{\max} \varepsilon_a E = (\sigma'_f)^2 (2N)^{2b} + \sigma'_f \varepsilon'_f E (2N)^{b+c} \quad (19)$$

where $\sigma_{\max} = \sigma_m + \sigma_a$ and ε_a is the alternating strain. If σ_{\max} is zero, (19) predicts infinite life, which implies that tension must be present for fatigue fractures to occur. Both (18) and (19) have been used to handle mean stress effect.

In Ong (1993), fatigue lives are calculated for 49 steels using published values of $\sigma'_f, \varepsilon'_f, b, c$. These lives were compared with lives calculated using some of the approximation methods described above. These include the original and modified versions of the four-point correlation method, the original universal slopes method, and the method of Mitchell et al. The lives covered a range from 10 to 10^7 reversals. The steels covered values of ultimate tensile strength from 345 MPa to 2,585 MPa, and Brinell hardness values from 80 to 660. They included the steels SAE1005, SAE1015, and SAE1045.

In all cases, correlation between the “experimental” and the “estimated” lives was poor. The modified four-point correlation method was found to be slightly better than the original universal slopes method and to be the best within the methods studied.

In another study conducted by Park and Song (1995), six such methods were evaluated and compared. These consisted of the universal slopes and four-point correlation methods by Manson (1965), the modified universal slopes method by Muralidharan and Manson (1988), the uniform material method by Baumel and Seeger (1990), the modified four-point correlation method by Ong (1993), and the method proposed by Mitchell et al. (1977). A total of 138 materials were used in the study, including unalloyed steels, low-alloy steels, high-alloy steels, aluminum alloys, and titanium alloys, with low-alloy steels providing the most data. Among the correlations compared, those proposed by Muralidharan and Manson (1988), Baumel and Seeger (1990), and Ong (1993) yielded good

predictions according to Park and Song (1995). The modified universal slopes method was concluded to provide the best correlation.

In the study carried out by Roessle and Fatemi (2000), they also compared their simple formula, which uses only hardness and modulus of elasticity for estimation of the strain-life curve, with the modified universal slopes method and found their simple formula results in somewhat better and more conservative predictions over the entire fatigue life regime. A similar study was carried out by Lee and Song (2006). They found that the (direct) hardness method proposed by Roessle and Fatemi (2000) provides excellent estimation results for steels. Lee and Song (2006) then proposed the indirect hardness methods that can be successfully applied to estimate fatigue properties for aluminum alloys and titanium alloys. The medians method proposed by Meggiolaro and Castro (2004) is found to provide the best estimation results for aluminum alloys. Based on the results obtained, some guidelines are provided for estimating fatigue properties from simple tensile data or hardness. In addition, a new relationship of ultimate tensile strength versus hardness is proposed for titanium alloys.

Most of the existing methods for estimating $\varepsilon \sim N$ parameters are based on a relatively limited amount of experimental data. In addition, sound statistical evaluations of the popular rules of thumb used in practice to estimate fatigue properties are scarce. Meggiolaro and Castro (2004) presented an extensive statistical evaluation of the existing Coffin–Manson parameter estimates based on monotonic tensile and uniaxial fatigue properties of 845 different metals, including 724 steels, 81 aluminum alloys, and 15 titanium alloys. From the collected data, it is shown that all correlations between the fatigue ductility coefficient ε'_f and the monotonic tensile properties are poor and that it is statistically sounder to estimate ε'_f based on constant values for each alloy family. Based on this result, a new estimation method that uses the medians of the individual parameters of the 845 materials is proposed.

Applications

The use of fatigue information in design has increased steadily over the years, and fatigue design philosophy has changed from one based on endurance limit approaches to one based upon a more precise assessment of fatigue durability. Consequently, modern methods of fatigue analysis are now utilized at the design stage. One such method is the local strain approach, which is now commonly adopted by most industries for fatigue analysis (Ong 1993). As explained above, the approach combines

the measure/design service loads imposed upon the structures at the most highly strained locations and the materials' fatigue properties (Ong 1993). The materials' fatigue properties are characterized by the strain-life curves, obtained from strain-controlled fatigue testing of smooth specimens or the engineering simplification methods introduced above. To apply the strain-based approach to an engineering component, such as a beam or a notched member, an analysis relating applied load and strain at the expected failure location is needed (Dowling 2007).

As strain-based approach can involve more detailed analysis of localized yielding that may occur at stress raisers during cyclic loading, it has been widely used in the design of the components of vehicles, machines, airplanes, and ship structures. Considering the cyclic properties of materials and the served environments, strain-based fatigue analysis has been favored in some other important industrial structures, such as nuclear power, thermo-electric, thermo-power, and thermo-mechanical plants and components (Zhao et al. 2008).

Although strain-based approaches are the improvement of the stress-based approaches under higher stress range, they are fundamentally based on the linear cumulative damage rule. The limitations mentioned in the entry on ► [damage accumulation](#) still exist in all these approaches and a way forward is to use the fatigue crack propagation theory.

Cross-References

- [Damage Accumulation](#)
- [Fatigue](#)
- [Fatigue Limit](#)

References

- A. Baumeel Jr., T. Seeger, *Materials Data for Cyclic Loading – Supplement I* (Elsevier Science, Amsterdam, 1990)
- C.C. Chu, Fatigue damage calculation using the critical plane approach. *J. Eng. Mater. Technol. ASME* **117**, 41–49 (1995)
- N.E. Dowling, *Mechanical Behavior of Materials-Engineering Methods for Deformation, Fracture, and Fatigue*, 3rd edn. (Pearson Prentice Hall, Upper Saddle River, 2007)
- K.S. Lee, J.H. Song, Estimation methods for strain-life fatigue properties from hardness. *Int. J. Fatigue* **28**, 386–400 (2006)
- S.S. Manson, Fatigue: a complex subject-some simple approximations. *Exp. Mech. J. Soc. Exp. Stress Anal.* **5**(7), 193–226 (1965)
- S.S. Manson, M.H. Hirschberg, *Fatigue: An Interdisciplinary Approach* (Syracuse University Press, Syracuse, 1964), p. 133
- M.A. Meggiolaro, J.T.P. Castro, Statistical evaluation of strain-life fatigue crack initiation predictions. *Int. J. Fatigue* **26**, 463–476 (2004)
- M. R. Mitchell, D. F. Spie, E. M. Caulfeld, *Fundamentals of Modern Fatigue Analysis, Fracture Control Program Report No. 26*, University of Illinois, USA, 1977, pp. 385–410
- U. Muralidharan, S.S. Manson, Modified universal slopes equation for estimation of fatigue characteristics. *J. Eng. Mater. Technol. Trans. Am. Soc. Mech. Eng.* **110**, 55–58 (1988)
- J.H. Ong, An improved technique for the prediction of axial fatigue life from tensile data. *Int. J. Fatigue* **15**(3), 213–219 (1993)
- J.H. Park, J.H. Song, Detailed evaluation of methods for estimation of fatigue properties. *Int. J. Fatigue* **17**(5), 365–373 (1995)
- D.T. Raske, J. Morrow, Mechanics of materials in low cycle fatigue testing, in *Manual on Low Cycle Fatigue Testing. ASTM STP 465* (American Society for Testing and Materials, Philadelphia, 1969), pp. 1–825
- M.L. Roessle, A. Fatemi, Strain-controlled fatigue properties of steels and some simple approximations. *Int. J. Fatigue* **22**, 495–511 (2000)
- K.N. Smith, P. Watson, T.H. Topper, A stress-strain function for the fatigue of metals. *J. Mater. ASTM* **5**(4), 767–778 (1970)
- D.F. Socie, G.B. Marquis, *Multiaxial Fatigue* (SAE, Warrendale, 2000)
- Y.X. Zhao, B. Yang, Z.Y. Zhai, The framework for a strain-based fatigue reliability analysis. *Int. J. Fatigue* **30**, 493–501 (2008)

Stress Concentration

NORIO HASEBE

Department of Civil Engineering, Nagoya Institute of Technology, Nagoya, Japan

Synonyms

[Stress concentration factor \(SCF\)](#); [Stress concentration value \(SCV\)](#); [Stress intensity factor \(SIF\)](#)

Definition

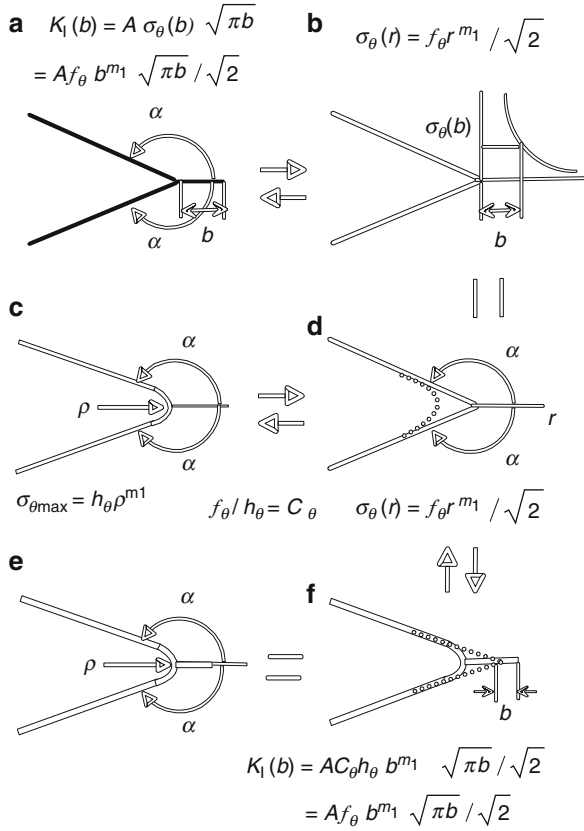
A notch is defined by two intersecting lines representing surfaces free of stress and linear elastic material located in the obtuse, notch angle denoted by 2α and the radius of curvature at the notch tip by ρ (Fig. 1c). For each notch of plane stress and thin plate bending problems, the relationships among sharp, round, and cracked notches using notch factors, 2α and ρ , are stated. Free and fixed boundary conditions are considered for each problem (Hasebe and Iida 1990).

Scientific Fundamentals

Sharp Notch and the Intensity of Corner for Plane Elastic Stress State

Sharp Notch with Free Boundary for Symmetric Stress State

Using the polar coordinates in Fig. 1d, the stress components near the sharp corner (V-shaped notch) are expressed as follows (Williams 1952a):



Stress Concentration, Fig. 1 Relations among round, sharp, and cracked notches

$$\begin{aligned} \sigma_r &= - \sum (m_j + 1) r^{m_j} \{ B_j (m_j - 2) \cos m_j \theta \\ &\quad + D_j (m_j + 2) \cos(m_j + 2) \theta \} \\ \sigma_\theta &= \sum (m_j + 1) (m_j + 2) r^{m_j} \{ B_j \cos m_j \theta \\ &\quad + D_j \cos(m_j + 2) \theta \} \\ \tau_{r\theta} &= \sum (m_j + 1) r^{m_j} \{ B_j m_j \sin m_j \theta \\ &\quad + D_j (m_j + 2) \sin(m_j + 2) \theta \} \end{aligned} \quad (1)$$

where $B_j \cos m_j \alpha + D_j \cos(m_j + 2) \alpha = 0$ holds, and $m_j (j = 1, 2, 3, \dots)$ are the roots of the following equation:

$$(m + 1) \sin 2\alpha + \sin\{2(m + 1)\alpha\} = 0 \quad (2)$$

The roots $m_j (j = 1, 2, 3, \dots)$ depend on 2α . The values are listed in Table 1. The first root m_1 is $0 \geq m_1 \geq -0.5$ for every angle of $2\alpha \geq 180^\circ$ and the stress component σ_θ symmetric to the bisector of the corner angle ($\theta = 0$) is expressed as

Stress Concentration, Table 1 Values of m_1 and m_2 of (2)

$2\alpha (^\circ)$	m_1	m_2
180	0.0	1.00000
190	-0.09996	1.00180
200	-0.18130	1.01826
210	-0.24803	$1.10629 \pm i0.09610$
220	-0.30284	$1.00565 \pm i0.19838$
230	-0.34773	$0.91527 \pm i0.23695$
240	-0.38427	$0.83355 \pm i0.25225$
250	-0.41372	$0.75925 \pm i0.25400$
260	-0.43716	$0.69141 \pm i0.24634$
270	-0.45552	$0.62926 \pm i0.23125$
280	-0.46960	$0.57214 \pm i0.20945$
290	-0.48015	$0.51955 \pm i0.18048$
300	-0.48778	$0.47103 \pm i0.14185$
310	-0.49307	$0.42623 \pm i0.08316$
320	-0.49651	0.46701
330	-0.49855	0.20296
340	-0.49957	0.12541
350	-0.49995	0.05884
360	-0.50000	0.00000

$$\begin{aligned} \sigma_\theta(r) &= \sum (m_j + 1)(m_j + 2) r^{m_j} (B_j + D_j) \\ &\equiv f_\theta r^{m_1} / \sqrt{2} + f_{\theta 2} r^{m_2} + f_{\theta 3} r^{m_3} + \dots \end{aligned} \quad (3)$$

The coefficient f_θ of the first term is defined as “the intensity of corner.” The factor $\sqrt{2}$ is given for the corresponding term to the stress intensity factor (SIF).

When the notch angle $2\alpha = 360^\circ$ corresponding to a crack, the component σ_θ is expressed as follows:

$$\sigma_\theta(r) = K_I / \sqrt{2\pi r} + f_2 + f_3 \sqrt{r} + \dots \quad (4)$$

K_I is defined as mode I stress intensity factor (SIF). Therefore, the intensity of corner f_θ corresponds to SIF in the case of a crack and the expansion for the notch with an arbitrary angle (Hasebe and Iida 1983).

Sharp Notch with Fixed Boundary for Symmetric Stress State

The stress components near the tip of sharp corner for fixed boundary are (Williams 1952a)

$$\begin{aligned}
\sigma_r &= - \sum (m_j + 1) r^{m_j} [\{(m_j + 2)A_j - 2C_j\} \\
&\quad \times \cos m_j \theta + C_j m_j \cos(m_j - 2)\theta] \\
\sigma_\theta &= \sum (m_j + 1) r^{m_j} [\{(m_j + 2)A_j + 2C_j\} \\
&\quad \times \cos m_j \theta + C_j m_j \cos(m_j - 2)\theta] \\
\tau_{r\theta} &= \sum (m_j + 1) r^{m_j} [\{(m_j + 2)A_j \sin m_j \theta \\
&\quad + C_j m_j \sin(m_j - 2)\theta\}]
\end{aligned} \quad (5)$$

where $A_j(m_j + 2) \sin(m_j + 2)\alpha + C_j(m_j + v + 1) \sin m_j \alpha = 0$ holds, and A_j and C_j are determined by loading condition and geometrical shape. The roots $m_j (j = 1, 2, 3, \dots)$ are those of the following equation:

$$\kappa \sin\{(m + 1)2\alpha\} - (m + 1) \sin 2\alpha = 0 \quad (6)$$

where $\kappa = (3 - \nu)/(1 + \nu)$, ν being Poisson's ratio. The values of $m_j (j = 1, 2)$ are listed in Table 2. In the symmetric axis ($\theta = 0$), σ_r is expressed by

$$\begin{aligned}
\sigma_r &= - \sum (m_j + 1) r^{m_j} \{(m_j + 2)A_j + (m_j - 2)C_j\} \\
&\equiv f_r r^{m_1} + f_{r2} r^{m_2} + f_{r3} r^{m_3} + \dots
\end{aligned} \quad (7)$$

where f_r is defined as “the intensity of the corner.” Because σ_r is the normal component of stress, f_r represents the intensity for the debonding (Iida et al. 1987a).

Sharp Notch with Fixed Boundary for Asymmetric Stress State

The stress components near the sharp corner for asymmetric stress state to the bisector of the sharp corner ($\theta = 0$) are expressed as follows:

$$\begin{aligned}
\sigma_r &= - \sum (m_j + 1) r^{m_j} [\{(m_j + 2)B_j - 2D_j\} \sin m_j \theta \\
&\quad + D_j m_j \sin(m_j - 2)\theta] \\
\sigma_\theta &= \sum (m_j + 1) r^{m_j} [\{(m_j + 2)B_j + 2D_j\} \sin m_j \theta \\
&\quad + D_j m_j \sin(m_j - 2)\theta] \\
\tau_{r\theta} &= - \sum (m_j + 1) r^{m_j} [\{(m_j + 2)B_j \cos m_j \theta \\
&\quad + D_j m_j \cos(m_j - 2)\theta\}]
\end{aligned} \quad (8)$$

where

$$B_j(m_j + 2) \sin(m_j + 2)\alpha + D_j(\kappa - m_j - 1) \sin m_j \alpha = 0 \quad (9)$$

In (9), B_j and D_j are determined by the loading condition, and $m_j (j = 1, 2, 3, \dots)$ are the roots of the following equation:

$$\kappa \sin\{2(m + 1)\alpha\} + (m + 1) \sin 2\alpha = 0 \quad (10)$$

and the values are listed in Table 3.

Stress Concentration, Table 2 Values of m_1 and m_2 in (6)

2α (°)	$\kappa = 1.0$		$\kappa = 5/3$		$\kappa = 2$		$\kappa = 3$	
	m_1	m_2	m_1	m_2	m_1	m_2	m_1	m_2
180	0.00000	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000	1.00000
190	0.00000	0.79893	-0.02185	0.83667	-0.02713	0.84619	-0.03579	0.86217
200	0.00000	0.63053	-0.04340	0.69804	0.05342	0.71474	-0.06955	0.74270
210	0.00000	0.48581	-0.06531	0.57959	-0.07950	0.60187	-0.10186	0.63881
220	0.00000	0.35950	-0.08824	0.47836	-0.10591	0.50489	-0.13317	0.54830
230	0.00000	0.24804	-0.11275	0.39224	-0.13307	0.42176	-0.16379	0.46940
240	0.00000	0.14891	-0.13931	0.31962	-0.16127	0.35084	-0.19394	0.40063
250	0.00000	0.06022	-0.16812	0.25909	-0.19059	0.29068	-0.22371	0.34068
260	-0.01953	0.00000	-0.19905	0.20922	-0.22093	0.23989	-0.25311	0.28838
270	-0.09147	0.00000	-0.23166	0.16847	-0.25199	0.19716	-0.28205	0.24270
280	-0.15656	0.00000	-0.26524	0.13525	-0.28334	0.16117	-0.31042	0.20267
290	-0.21556	0.00000	-0.29904	0.10806	-0.31452	0.13072	-0.33805	0.16742
300	-0.26910	0.00000	-0.33234	0.08554	-0.34509	0.10474	-0.36478	0.13618
310	-0.31771	0.00000	-0.36548	0.06659	-0.37465	0.08229	-0.39047	0.10826
320	-0.36182	0.00000	-0.39536	0.05032	-0.40293	0.06260	-0.41500	0.08307
330	-0.40181	0.00000	-0.42441	0.03604	-0.42971	0.04502	-0.43827	0.06007
340	-0.43799	0.00000	-0.45158	0.02318	-0.45486	0.02901	-0.46021	0.03882
350	-0.47065	0.00000	-0.47679	0.01129	-0.47830	0.01414	-0.48079	0.01891
360	-0.50000	0.00000	-0.50000	0.00000	-0.50000	0.00000	-0.50000	0.00000

Stress Concentration, Table 3 Values of m_1 and m_2 in (10)

2α (°)	$\kappa = 1$		$\kappa = 5/3$		$\kappa = 2$		$\kappa = 3$	
	m_1	m_2	m_1	m_2	m_1	m_2	m_1	m_2
180	0.00000	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000	1.00000
190	−0.09996	1.00180	−0.08153	0.95665	−0.07683	0.94593	−0.06889	0.92847
200	−0.18130	1.01826	−0.15022	0.91575	−0.14218	0.89452	−0.12851	0.86126
210	−0.24803	1.10629 ± i0.09610	−0.20829	0.87753	−0.19794	0.84509	−0.18025	0.79729
220	−0.30284	1.00565 ± i0.19838	−0.25746	0.84215	−0.24560	0.79669	−0.22525	0.73560
230	−0.34773	0.91527 ± i0.02369	−0.29910	0.80998	−0.28636	0.74795	−0.26446	0.67534
240	−0.38427	0.83355 ± i0.25225	−0.33432	0.78282	−0.32122	0.69706	−0.29867	0.61585
250	−0.41372	0.75915 ± i0.25400	−0.36405	0.77037 ± i0.02624	−0.35101	0.64194	−0.32856	0.55674
260	−0.43716	0.69141 ± i0.24634	−0.38908	0.58129 ± i0.03724	−0.37646	0.58129	−0.35473	0.49790
270	−0.45552	0.62926 ± i0.23125	−0.41011	0.61233	−0.39819	0.51586	−0.37768	0.43954
280	−0.46960	0.57214 ± i0.20945	−0.42773	0.51174	−0.41675	0.44817	−0.39786	0.38207
290	−0.48015	0.51955 ± i0.18048	−0.44249	0.42589	−0.43262	0.38087	−0.41568	0.32600
300	−0.48778	0.47103 ± i0.14185	−0.45485	0.34824	−0.44625	0.31578	−0.43149	0.27180
310	−0.49307	0.42623 ± i0.08316	−0.46525	0.27725	−0.45801	0.25388	−0.44561	0.21983
320	−0.49651	0.30269	−0.47407	0.21210	−0.46825	0.19562	−0.45833	0.17039
330	−0.49855	0.20296	−0.48166	0.15219	−0.47730	0.14114	−0.46991	0.12362
340	−0.49957	0.12541	−0.48832	0.09709	−0.48544	0.09043	−0.48057	0.07962
350	−0.49995	0.05884	−0.49434	0.04646	−0.49292	0.04342	−0.49054	0.03842
360	−0.50000	0.00000	−0.50000	0.00000	−0.50000	0.00000	−0.50000	0.00000

The stress component $\tau_{r\theta}$ along the symmetric axis ($\theta = 0$) is expressed as

$$\tau_{r\theta} = - \sum (m_j + 1) r^{m_j} \{ (m_j + 2) B_j + D_j m_j \} \quad (11)$$

$$\equiv f_{r0} r^{m_1} + f_{r02} r^{m_2} + f_{r03} r^{m_3} + \dots$$

where $f_{r\theta}$ is also defined as “the intensity of the corner.” This $f_{r\theta}$ expresses the intensity for slipping in a fixed edge.

Round Notch and the Stress Concentration for Plane Elastic Stress State

The stress concentration value (SCV) at the tip of the round notch with free boundary and fixed boundary is generally expressed in terms of the radius of curvature ρ (Hasebe 1971; Hasebe et al. 1986, 1987)

$$SCV = \sum h_j \rho^{m_j} \quad (12)$$

where h_j is determined by loading condition and geometric shape.

Free Boundary for Symmetric Stress State

The maximum tangential stress $\sigma_{\theta \max}$ at the notch tip is expressed by (12) as follows:

$$\sigma_{\theta \max} = h_0 \rho^{m_1} + h_{02} \rho^{m_2} + h_{03} \rho^{m_3} + \dots \quad (13)$$

Because the equation above converges quickly, SCV is obtained by the first two or three terms with sufficient accuracy (Hasebe and Kutanda 1978).

Fixed Boundary for Symmetric Stress State

Let the normal component, $\sigma_{r \max}$, at the notch tip be expressed by (12) as follows:

$$\sigma_{r \max} = h_r \rho^{m_1} + h_{r2} \rho^{m_2} + h_{r3} \rho^{m_3} + \dots \quad (14)$$

Because the equation above converges quickly, SCV is obtained by the first two or three terms with sufficient accuracy.

The following relationship between the tangential and normal components of stresses, σ_{θ} and σ_r , holds on the fixed boundary (Hasebe 1979):

$$\sigma_{\theta} / \sigma_r = (3 - \kappa) / (1 + \kappa). \quad (15)$$

Fixed Boundary for Asymmetric Stress State

The maximum shear stress $\tau_{r\theta\max}$ at the notch tip is expressed by (12) as

$$\tau_{r\theta\max} = h_{r\theta}\rho^{m_1} + h_{r\theta 2}\rho^{m_2} + h_{r\theta 3}\rho^{m_3} + \dots \quad (16)$$

Because the equation above converges quickly, SCV is obtained by the first two or three terms with sufficient accuracy.

Relationship Between the Intensity of the Corner and the Stress Concentration

Free Boundary for Symmetric Stress State

The relation between f_θ and h_θ in (3) and (13) is expressed as follows:

$$f_\theta/h_\theta = C_\theta \quad (17)$$

The equation above depends on only the notch angle without regard to loading condition. The values of C_θ are listed in Table 4. Therefore, if either f_θ or h_θ is known, another value can be obtained. SIF is related to SCV as follows:

$$K_I = \lim_{\rho \rightarrow 0} \frac{1}{2} \sigma_{\max} \sqrt{\pi \rho} \quad (18)$$

For the crack, therefore, K_I is expressed by h_θ as

$$K_I = \frac{1}{2} \sqrt{\pi} h_\theta \quad (19)$$

When some SCV and ρ are known, an expression of stress concentration can be formed by using the first two or three terms in (13). Using this expression, SCV can be calculated for any ρ . If the expression of SCV or $\sigma_{\theta\max}$ are known as the function of ρ , then f_θ is obtained from the following expression:

$$f_\theta = \lim_{\rho \rightarrow 0} C_\theta (\rho^{-m_1} \sigma_{\theta\max}) \quad (20)$$

Stress Concentration, Table 4 Values of C_θ in (17)

2α (°)	C_θ	2α (°)	C_θ
180	1.414	280	0.506
190	0.952	290	0.503
200	0.775	300	0.502
210	0.676	310	0.501
220	0.615	320	0.500
230	0.576	330	0.500
240	0.549	340	0.500
250	0.531	350	0.500
260	0.516	360	0.5
270	0.512		

Thus, for example, if SCV for small ρ is known, then h_θ is obtained by (13) and f_θ by (17). Accordingly, B_1 and C_1 in (1) are known and the stress distributions near the sharp corner are obtained; the converse is also true.

Fixed Boundary for Symmetric Stress State

The intensity of corner f_r in (7) is related to h_r in (14) as follows:

$$f_r/h_r = C_r \quad (21)$$

Equation 21 depends on the corner angle and Poisson's ratio without regard to loading condition. Table 5 contains the values of C_r . When $\sigma_{r\max}$ is known as the function of ρ , f_r is obtained in the same way as (20). The stress components along the boundary are expressed in the same way as the normal one;

$$\begin{aligned} \sigma_\theta &= f_\theta r^{m_1} + f_{\theta 2} r^{m_2} + f_{\theta 3} r^{m_3} + \dots \\ \sigma_{\theta\max} &= h_\theta \rho^{m_1} + h_{\theta 2} \rho^{m_2} + h_{\theta 3} \rho^{m_3} + \dots \end{aligned} \quad (22)$$

Stress Concentration, Table 5 Values of C_r in (21)

2α (°)	$\kappa = 5/3$	$\kappa = 2$	$\kappa = 3$
	C_r	C_r	C_r
180	1.000	1.000	1.000
190	0.994	0.930	0.901
200	0.975	0.898	0.856
210	0.953	0.881	0.828
220	0.927	0.871	0.810
230	0.905	0.866	0.798
240	0.888	0.861	0.793
250	0.879	0.859	0.790
260	0.877	0.860	0.791
270	0.883	0.863	0.796
280	0.893	0.871	0.804
290	0.906	0.880	0.814
300	0.922	0.894	0.825
310	0.941	0.912	0.840
320	0.966	0.934	0.861
330	0.999	0.963	0.884
340	1.042	1.004	0.918
350	1.108	1.063	0.966
360	1.238	1.178	1.060

From (15) for small ρ

$$h_\theta/h_r \approx \sigma_{\theta \max}/\sigma_{r \max} = (3 - \kappa)/(1 + \kappa) \quad (23)$$

and from (5), (7), and (22), the following equation is derived:

$$f_\theta/h_r = - \left[\begin{array}{l} \{(m_1 + \kappa + 1) \sin(m_1 \alpha)\} \\ -(m_1 + 2) \sin\{(m_1 + 2)\alpha\} \end{array} \right] / \left[\begin{array}{l} \{(m_1 + \kappa + 1) \sin(m_1 \alpha)\} \\ -(m_1 - 2) \sin\{(m_1 + 2)\alpha\} \end{array} \right] \quad (24)$$

If any one of $f_\theta, h_\theta, f_r, h_r$ are known, other values can be obtained and then SCV and the stress components near the corner can be calculated (Hasebe et al. 1987).

Fixed Boundary for Asymmetric Stress State

The intensity of corner $f_{r\theta}$ in (11) is related to $h_{r\theta}$ in (16) as follows:

$$f_{r\theta}/h_{r\theta} = C_{r\theta} \quad (25)$$

$C_{r\theta}$ depends on the corner angle and Poisson's ratio without regard to loading condition, and are shown in Table 6.

When $\tau_{r\theta \max}$ is known as a function of ρ , $f_{r\theta}$ is obtained. If any one of $f_{r\theta}$ and $h_{r\theta}$ is known, then SCV and the stresses components near the corner can be obtained in the same way as the free boundary.

Sharp Notch and the Intensity of Corner for Thin Plate Bending Problem

Sharp Notch with Free Boundary for Symmetric Stress State

For the sharp V-shaped notch, the bending and torsional moments near the tip are expressed as follows (Williams 1952b):

$$\begin{aligned} M_r &= - \sum D r^{m_j} \\ &\quad \times \left[\begin{array}{l} \{(m_j + 1)(m_j + 2) + v(m_j + 2) - v m_j^2\} F_j \cos m_j \theta \\ + (m_j + 2)\{(m_j + 1) + v - v(m_j + 2)\} H_j \cos(m_j + 2)\theta \end{array} \right] \\ M_\theta &= - \sum D r^{m_j} \\ &\quad \times \left[\begin{array}{l} \{(m_j + 2) + v(m_j + 1)(m_j + 2) - m_j^2\} F_j \cos m_j \theta \\ + (m_j + 2)\{1 + v(m_j + 1) - (m_j + 2)\} H_j \cos(m_j + 2)\theta \end{array} \right] \\ H_{r\theta} &= \sum D r^{m_j} (m_j + 1) [(1 - v) m_j F_j \sin m_j \theta \\ &\quad + (1 - v)(m_j + 2) H_j \sin(m_j + 2)\theta] \end{aligned} \quad (26)$$

Stress Concentration, Table 6 Values of $C_{r\theta}$ in (25)

2α (°)	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$
	$C_{r\theta}$	$C_{r\theta}$	$C_{r\theta}$
180	1.000	1.000	1.000
190	0.529	0.687	0.760
200	0.388	0.583	0.648
210	0.316	0.511	0.575
220	0.270	0.458	0.523
230	0.236	0.414	0.483
240	0.202	0.379	0.452
250	0.172	0.351	0.427
260	0.142	0.326	0.405
270	0.116	0.305	0.388
280	0.092	0.286	0.373
290	0.071	0.272	0.362
300	0.053	0.259	0.352
310	0.037	0.249	0.345
320	0.026	0.241	0.341
330	0.016	0.235	0.338
340	0.009	0.232	0.339
350	0.004	0.233	0.344
360	0.0	0.236	0.354

where

$$F_j(4 + m_j - v m_j) \sin m_j \alpha + H_j(1 - v)(m_j + 2) \sin(m_j + 2)\alpha = 0 \quad (27)$$

holds; D is the flexural rigidity and F_j and H_j are determined by the loading condition and shape.

The values m_j ($j = 1, 2, 3, \dots$) are roots of the following equation:

$$(3 + v) \sin\{2(m + 1)\alpha\} - (1 - v)(m + 1) \sin 2\alpha = 0 \quad (28)$$

and depend on 2α and Poisson's ratio. Table 7 shows the first two roots of (28). The bending moment M_θ in the symmetric axis is expressed as follows:

$$\begin{aligned} M_\theta &= - \sum D r^{m_j} \left[\begin{array}{l} \{(m_j + 2) + v(m_j + 1)\} F_j \\ \times (m_j + 2) - m_j^2 \end{array} \right] \\ &\quad + (m_j + 2) \left[\begin{array}{l} 1 + v(m_j + 1) \\ - (m_j + 2) \end{array} \right] H_j \\ &\equiv f_\theta r^{m_1} / \sqrt{2} + f_{\theta 2} r^{m_2} + f_{\theta 3} r^{m_3} + \dots \end{aligned} \quad (29)$$

Stress Concentration, Table 7 Values m_1 and m_2 in (28)

2α (°)	$\nu = 0.0$		$\nu = 0.25$		$\nu = 0.5$	
	m_1	m_2	m_1	m_2	m_1	m_2
180	0.00000	1.00000	0.00000	1.00000	0.00000	1.00000
190	-0.03579	0.86217	-0.04104	0.87209	-0.04549	0.88066
200	-0.06955	0.74270	-0.07916	0.76007	-0.08722	0.77512
210	-0.10186	0.63881	-0.11495	0.66165	-0.12581	0.68144
220	-0.13316	0.54830	-0.14885	0.57495	-0.16175	0.59801
230	-0.16379	0.46940	-0.18121	0.49843	-0.19543	0.52349
240	-0.19394	0.40063	-0.21226	0.43077	-0.22714	0.45674
250	-0.22371	0.34068	-0.24217	0.37083	-0.25711	0.39674
260	-0.25311	0.28839	-0.27102	0.31760	-0.28550	0.34274
270	-0.28205	0.24270	-0.29885	0.27020	-0.31246	0.29388
280	-0.31042	0.20267	-0.32566	0.22785	-0.33806	0.24955
290	-0.33804	0.16743	-0.35143	0.18982	-0.36238	0.20917
300	-0.36478	0.13618	-0.37613	0.15549	-0.38547	0.17220
310	-0.39047	0.10826	-0.39970	0.12431	-0.40736	0.13821
320	-0.41500	0.08307	-0.42213	0.09577	-0.42809	0.10677
330	-0.43827	0.06007	-0.44339	0.06943	-0.44769	0.07753
340	-0.46021	0.03882	-0.46345	0.04491	-0.46619	0.05017
350	-0.48079	0.01892	-0.48232	0.02187	-0.48362	0.02441
360	-0.50000	0.00000	-0.50000	0.00000	-0.50000	0.00000

where f_θ is also defined as “the intensity of corner.” A method for obtaining f_θ was described in Hasebe and Iida (1983).

Sharp Notch with Fixed Boundary for Symmetric Stress State

The bending and torsional moments near the sharp corner with fixed boundary are expressed as follows:

$$\begin{aligned}
 M_r &= - \sum D r^{m_j} \\
 &\quad \times \left[\left\{ (m_j + 1)(m_j + 2) + \nu(m_j + 2) - \nu m_j^2 \right\} E_j \cos m_j \theta \right. \\
 &\quad \left. + (m_j + 2) \{ (m_j + 1) + \nu - \nu(m_j + 2) \} G_j \cos(m_j + 2) \theta \right] \\
 M_\theta &= - \sum D r^{m_j} \\
 &\quad \times \left[\left\{ (m_j + 2) + \nu(m_j + 1)(m_j + 2) - m_j^2 \right\} E_j \cos m_j \theta \right. \\
 &\quad \left. + (m_j + 2) \{ 1 + \nu(m_j + 1) - (m_j + 2) \} G_j \cos(m_j + 2) \theta \right] \\
 H_{r\theta} &= \sum D r^{m_j} (m_j + 1) [(1 - \nu) m_j E_j \sin m_j \theta \\
 &\quad + (1 - \nu)(m_j + 2) G_j \sin(m_j + 2) \theta]
 \end{aligned}
 \tag{30}$$

where

$$E_j \cos m_j \alpha + G_j \cos(m_j + 2) \alpha = 0 \tag{31}$$

holds; E_j and G_j are determined by the loading condition and shape and $m_j (j = 1, 2, 3, \dots)$ are roots of the following equation:

$$(m + 1) \sin(2\alpha) + \sin\{2(m + 1)\alpha\} = 0 \tag{32}$$

Because (32) is the same as (2), the values of m_1 and m_2 are given in Table 1. From (30), M_r and M_θ in the symmetric axis are expressed as follows:

$$\begin{aligned}
 M_r &= - \sum D r^{m_j} \left[\left\{ (m_j + 1)(m_j + 2) + \nu(m_j + 2) - \nu m_j^2 \right\} E_j \right. \\
 &\quad \left. + (m_j + 2) \{ (m_j + 1) + \nu - \nu(m_j + 2) \} G_j \right] \\
 &\equiv f_r r^{m_1} + f_{r2} r^{m_2} + f_{r3} r^{m_3} + \dots \\
 M_\theta &= - \sum D r^{m_j} \left[\left\{ (m_j + 2) + \nu(m_j + 1)(m_j + 2) - m_j^2 \right\} E_j \right. \\
 &\quad \left. + (m_j + 2) \{ 1 + \nu(m_j + 1) - (m_j + 2) \} G_j \right] \\
 &\equiv f_\theta r^{m_1} + f_{\theta 2} r^{m_2} + f_{\theta 3} r^{m_3} + \dots
 \end{aligned}
 \tag{33}$$

where f_r and f_θ are defined as “the intensity of corner.” These f_r and f_θ represent the intensity relating to the debonding in the fixed edge and to the crack initiation into the elastic plate, respectively.

Round Notch and the Stress Concentration for Thin Plate Bending Problem

SCV of bending and torsional moment in the bisector of notch is expressed by

$$SCV = \sum h_j \rho^{m_j} \quad (34)$$

Free Boundary Condition

The bending moment in the tangential direction, $M_{\theta \max}$, is expressed from (34) as follows:

$$M_{\theta \max} = h_0 \rho^{m_1} + h_{02} \rho^{m_2} + h_{03} \rho^{m_3} + \dots \quad (35)$$

where $m_j (j = 1, 2, 3, \dots)$ are roots of (28), and h_0 is determined by loading and geometrical shape. Because (35) converges fast, SCV can be expressed by the first two or three terms with sufficient accuracy (Hasebe 1971).

Fixed Boundary Condition

The normal and tangential moment $M_{r \max}$ and $M_{\theta \max}$ at the corner are expressed as follows (Hasebe et al. 1986):

$$\begin{aligned} M_{r \max} &= h_r \rho^{m_1} + h_{r2} \rho^{m_2} + h_{r3} \rho^{m_3} + \dots \\ M_{\theta \max} &= h_{\theta} \rho^{m_1} + h_{\theta 2} \rho^{m_2} + h_{\theta 3} \rho^{m_3} + \dots \end{aligned} \quad (36)$$

where $m_j (j = 1, 2, 3, \dots)$ are roots of (32), and h_j is determined by loading and geometrical shape. Because (36) converges fast, SCV can be expressed by the first two or three terms with sufficient accuracy. The normal moment M_n is related to the tangential moment M_t , and the torsional moment H_{nt} is equal to zero in the fixed edge:

$$M_t = \nu M_n, \quad H_{nt} = 0 \quad (37)$$

Relationship Between the Intensity of Corner and Stress Concentration

Free Boundary for Symmetric Stress State

The intensity of corner f_{θ} in (29) is related to h_{θ} in (35) as follows:

$$f_{\theta}/h_{\theta} = C_{\theta}(3 + \nu)/(1 + \nu) \quad (38)$$

Equation 38 depends on the corner angle and Poisson's ratio without regard to loading condition. Table 8 contains the value C_{θ} . When some values of ρ and $M_{\theta \max}$ are known, the expression of $M_{\theta \max}$ is formed from (35). If $M_{\theta \max}$ is known as the function of ρ , then f_{θ} is obtained by the following expression:

$$f_{\theta} = \{C_{\theta}(3 + \nu)/(1 + \nu)\} \lim_{\rho \rightarrow 0} (\rho^{-m_1} M_{\theta \max}) \quad (39)$$

For the crack, C_{θ} is 0.5 and f_{θ} is the stress concentration factor for the crack. If any one of the factors of the

Stress Concentration, Table 8 Values of C_{θ} in (38)

2α (°)	$\nu = 0.0$	$\nu = 0.25$	$\nu = 0.5$
	C_{θ}	C_{θ}	C_{θ}
180	0.471	0.544	0.606
190	0.425	0.478	0.522
200	0.403	0.448	0.483
210	0.390	0.428	0.458
220	0.382	0.415	0.440
230	0.376	0.407	0.428
240	0.373	0.400	0.420
250	0.372	0.397	0.414
260	0.373	0.394	0.410
270	0.375	0.394	0.408
280	0.379	0.395	0.407
290	0.384	0.397	0.408
300	0.389	0.402	0.411
310	0.396	0.406	0.415
320	0.406	0.414	0.422
330	0.417	0.424	0.430
340	0.433	0.438	0.443
350	0.456	0.458	0.462
360	0.5	0.5	0.5

notch is known, then SCV and the components of moment at the corner tip are obtained in the same way as the plane problem.

Fixed Boundary

The intensity f_r in (33) is related to h_r in (36) and f_{θ} to h_{θ} as follows:

$$f_r/h_r = C_r, \quad f_{\theta}/h_{\theta} = C_{\theta} \quad (40)$$

Equation 40 depends on the corner angle and Poisson's ratio without regard to loading condition.

Table 9 contains the value C_r . From (36) and (37), for small ρ , $h_{\theta}/h_r \approx M_{\theta \max}/M_{r \max} = \nu$ and from (30) and (33), the following equation is obtained:

$$\frac{f_{\theta}}{f_r} = \frac{\left[\begin{aligned} &\{-(m_1 - 2) + \nu(m_1 + 2)\} \sin\{(m_1 + 2)\alpha\} \\ &+ (1 - \nu)m_1 \sin(m_1\alpha) \end{aligned} \right]}{\left[\begin{aligned} &\{(m_1 + 2) - \nu(m_1 - 2)\} \sin\{(m_1 + 2)\alpha\} \\ &- (1 - \nu)m_1 \sin(m_1\alpha) \end{aligned} \right]} \quad (41)$$

Stress Concentration, Table 9 Values of C_r in (40a)

2α (°)	$\nu = 0.0$	$\nu = 0.25$	$\nu = 0.5$
	C_r	C_r	C_r
180	1.000	1.000	1.000
190	0.673	0.675	0.676
200	0.548	0.564	0.573
210	0.438	0.510	0.526
220	0.435	0.478	0.503
230	0.407	0.458	0.493
240	0.389	0.444	0.487
250	0.376	0.438	0.487
260	0.367	0.436	0.488
270	0.362	0.435	0.491
280	0.358	0.434	0.498
290	0.356	0.434	0.506
300	0.355	0.435	0.513
310	0.354	0.437	0.519
320	0.354	0.439	0.523
330	0.354	0.440	0.526
340	0.354	0.441	0.528
350	0.354	0.442	0.529
360	0.354	0.442	0.530

If any one of f_θ , h_θ , f_r , h_r are known, other values can be obtained and then SCV and the stress components near the corner tip can be calculated (Iida et al. 1987b).

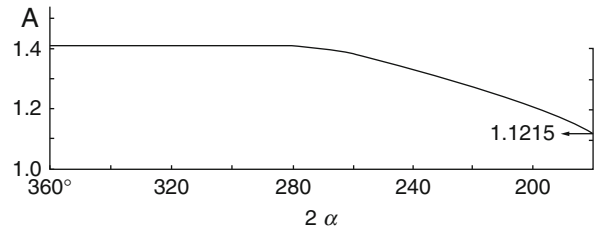
Relationship Between Crack and Notch Mechanics

Plane Elastic Problem for Free Boundary

As shown in Fig. 1a, a crack initiating from the V-shaped notch in the bisector is considered. SIF $K_I(b)$ with the crack length b is expressed by the $\sigma_\theta(b)$ in the bisector of the notch angle before crack initiation (Hasebe and Iida 1979):

$$K_I(b) = A\sigma_\theta(b)\sqrt{\pi b} \quad (42)$$

where A depends on the crack length. Though $\sigma_\theta(b)$ is not given in the form of a function, when stress values are known for some b , the expression of $\sigma_\theta(b)$ can be formed by using a few terms in (3). Using this expression of $\sigma_\theta(b)$, the value of K_I can be calculated. Near the notch tip (r or b is small), $\sigma_\theta(b)$ is expressed by only the first term in (3) and from (42):

Stress Concentration, Fig. 2 Values of A in (43)

$$\begin{aligned} \sigma_\theta(b) &= f_\theta b^{m_1} / \sqrt{2} \\ K_I(b) &= A f_\theta b^{m_1} \sqrt{\pi b} / \sqrt{2} \end{aligned} \quad (43)$$

For short cracks, the value A depends on the angle only and is shown in Fig. 2. From (43), it is known that SIF for a short crack is expressed by f_θ . Therefore if f_θ is known, the SIF for a short crack initiating from the notch can be calculated. Conversely, if the SIF for a short crack is known, the value f_θ can be computed from (43). As mentioned above, if f_θ is known, then arbitrary stress components near the notch tip and h_θ are obtained and SCV for small ρ is also obtained. These relations are shown in Fig. 1.

Thin Plate Bending Problem for Free Boundary

The SIF for thin plate bending, $K_B(b)$, is expressed by a crack length b by using $M_\theta(b)$ in the bisector of the notch angle before crack initiation as follows:

$$K_B(b) = \{(1 + \nu)/(3 + \nu)\} A M_\theta(b) \sqrt{b} \quad (44)$$

where A is shown in Fig. 3 and depends on notch angle, crack length, and Poisson's ratio. Because the bending moment is expressed by only the first term in (29) near the notch, the SIF tip, for a short crack, is obtained as follows:

$$K_B(b) = \{(1 + \nu)/(3 + \nu)\} A f_\theta b^{m_1} \sqrt{b} / \sqrt{2}. \quad (45)$$

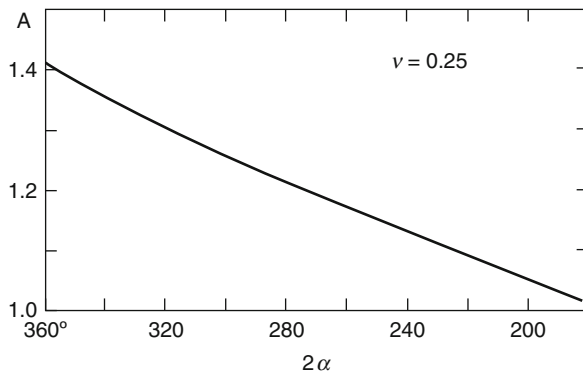
When f_θ is given, the SIF can be calculated from the equation above, and if the SIF for a short crack is given, then f_θ can be calculated by (45) and the stress components near the notch tip before crack initiation and SCV for small ρ are obtained (Hasebe and Iida 1990; Iida et al. 1990).

The stress concentration in the anti-plane shear problem is also investigated in the same way as the plane elastic and thin plate bending problems (Hasebe et al. 1987).

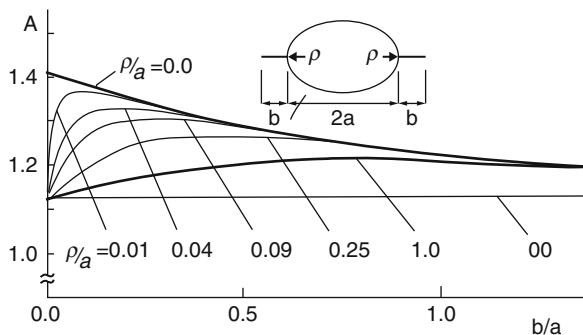
Key Applications

When the problem for two notches with different angles and subjected to different loadings is considered, the question is which notch is in more danger of cracking? It is judged by comparing the values of the SIF for the crack initiated in two notches. The SIF can be computed by (43), if the intensities of the corner are given. In order that a crack does not initiate from the notch, the intensity of the notch requires that the SIF for a suitable b in (43) must not exceed the fracture toughness value K_{Ic} .

The SIF of a crack initiating at the tip of the round notch can be given by (43) with the value A of Fig. 4, which shows the value A versus the ratio of crack length b and the semi-axis a of the elliptic hole for some ρ . In Fig. 3, when b is small enough, all values of A approach the value 1.1215 asymptotically.



Stress Concentration, Fig. 3 Values of A in (45)



Stress Concentration, Fig. 4 Values of A in (43)

Cross-References

► [Stress Intensity Factors](#)

References

- N. Hasebe, Stress analysis of a semi-infinite plate with a triangular notch or mound. *Proc. Jpn. Soc. Civil Eng.* **194**, 29 (1971)
- N. Hasebe, Uniform tension of a semi-infinite plate with a crack at an end of a stiffened edge. *Ing. Arch.* **48-2**, 129 (1979)
- N. Hasebe, J. Iida, A crack originating from a triangular notch on a rim of a semi-infinite plate under transverse bending. *Eng. Fract. Mech.* **11-4**, 645 (1979)
- N. Hasebe, J. Iida, Intensity of corner and stress concentration factor. *J. Eng. Mech. ASCE* **109-1**, 346 (1983)
- N. Hasebe, J. Iida, Notch mechanics for plane and thin plate bending problems. *Eng. Fract. Mech.* **37-1**, 87 (1990)
- N. Hasebe, Y. Kutanda, Calculation of stress intensity factor from stress concentration. *Eng. Fract. Mech.* **10-2**, 215 (1978)
- N. Hasebe, T. Sugimoto, T. Nakamura, Stress concentration in clamped edge of thin plate. *J. Eng. Mech. ASCE* **112-7**, 642 (1986)
- N. Hasebe, T. Sugimoto, T. Nakamura, Stress concentration of longitudinal shear problems. *J. Eng. Mech. ASCE* **113-9**, 1358 (1987)
- J. Iida, N. Hasebe, S. Matsura, Intensity of corner in fixed edge of Plane problem. *J. Eng. Mech. ASCE* **113-8**, 1194 (1987a)
- J. Iida, N. Hasebe, T. Nakamura, Intensity of corner in fixed edge of thin plate. *J. Eng. Mech. ASCE* **113-8**, 1138 (1987b)
- J. Iida, N. Hasebe, T. Nakamura, Approximate expressions for SIF of crack initiating from notch for thin plate bending and plane problems. *Eng. Fract. Mech.* **36-5**, 819 (1990)
- M.L. Williams, Stress singularities resulting from various boundary conditions in angular corner of plate in extension. *J. Appl. Mech.* **74**, 526 (1952a)
- M.L. Williams, Surface stress singularities resulting from various boundary conditions in angular corners of plates under bending, in *1st U.S. National Congress Applied Mechanics*, McGraw-Hill, New York, (1952b), p. 32

Stress Concentration Factor (SCF)

► [Stress Concentration](#)

Stress Concentration Value (SCV)

► [Stress Concentration](#)

Stress Intensity Factor (SIF)

► [Stress Concentration](#)

Stress Intensity Factors

ALAN T. ZEHNDER

Field of Theoretical and Applied Mechanics, Cornell University, Ithaca, NY, USA

Synonyms

K_I ; K_{II} ; K_{III} ; SIF

Definition

The stress intensity factor is the magnitude of the stress singularity at the tip of a mathematically sharp crack in a linear elastic material. Each mode of fracture has an associated stress intensity factor. The Mode-I, Mode-II, and Mode-III stress intensity factors, labeled K_I , K_{II} , and K_{III} are defined with respect to Fig. 1 by

$$K_I = \lim_{r \rightarrow 0} \sqrt{2\pi r} \sigma_{22}(r, \theta = 0)$$

$$K_{II} = \lim_{r \rightarrow 0} \sqrt{2\pi r} \sigma_{12}(r, \theta = 0)$$

$$K_{III} = \lim_{r \rightarrow 0} \sqrt{2\pi r} \sigma_{32}(r, \theta = 0)$$

The units of stress intensity factor are $[F/L^{3/2}]$. In MKS units stress intensity factors are typically given as $MPa\sqrt{m}$.

Scientific Fundamentals

Stress intensity factors arise from the solution of the problem of a two-dimensional crack in a homogeneous, isotropic, linearly elastic material. With respect to the coordinate system shown in Fig. 1, a crack is defined by surfaces at $\theta = \pm\pi$ across which a discontinuity in the displacement fields can occur. In the classical analysis of this problem it is assumed that the crack surfaces are traction free (i.e., no forces act of these surfaces). However cracks can and often do have contact forces and fluid pressure acting on the crack surfaces.

Using the idea of modes of fracture the problem of finding the stresses at a crack tip is broken down into two

two-dimensional problems. The σ_{11} , σ_{12} , and σ_{22} components of stress are calculated using the plane stress or plane strain theory of linear elasticity. The out-of-plane shear stresses σ_{13} and σ_{23} are calculated using the anti-plane shear theory of linear elasticity.

The simplest problem to solve is the Mode-III, or anti-plane shear crack. This solution demonstrates the underlying nature of the crack tip fields and allows the assumptions and limitations of fracture mechanics to be revealed. In anti-plane shear it is assumed that the out-of-plane displacement field u_3 is a function of the in-plane coordinate only and that the in-plane displacements are zero, i.e., $u_1 = u_2 = 0$, $u_3 = u(x_1, x_2)$. The resulting field equations for a static problem with no body forces are

$$\nabla^2 u = 0$$

$$\gamma_{\alpha 3} = \frac{1}{2} \frac{\partial u}{\partial x_\alpha}$$

$$\sigma_{\alpha 3} = \mu \frac{\partial u}{\partial x_\alpha},$$

where $\alpha = \{1, 2\}$, μ is the shear modulus, and $\gamma_{\alpha 3}$ are strain components. On the boundaries either the surface forces in the x_3 direction or the displacement in the x_3 direction can be prescribed, i.e.,

$$t_3 = \sigma_{\alpha 3} n_\alpha = t^*(x_1, x_2)$$

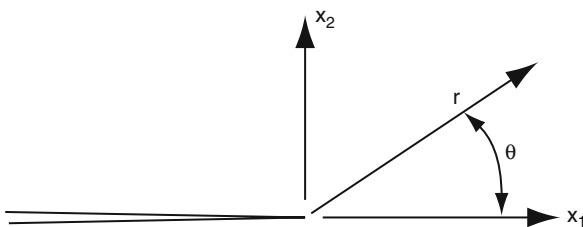
or

$$u = u^*(x_1, x_2),$$

where n_α are components of the surface normal vectors and t^* and u^* are prescribed functions on the boundary of the body under study. For the crack problem the only boundary conditions are $\sigma_{23} = 0$ on $\theta = \pm\pi$. The solution for the stress field can be expressed compactly using complex variables as

$$\sigma_{23} + i\sigma_{13} = \sum_{n=\dots, -3, -1, 1, 3, 5, \dots} A_n z^{n/2} \quad (1)$$

where $z = x_1 + ix_2$. The coefficients A_n are undetermined in this asymptotic analysis since only the traction-free crack surface boundary conditions are accounted for. To ensure that the crack tip displacements are finite, the stress can be no more singular than $z^{-1/2}$. The flaw in this argument is that the material will have some inelastic deformation at the crack tip and thus one cannot use the linear elastic stress field all the way to the crack tip. Modeling the crack inelastic zone as a hole or rigid inclusion in a linear elastic, isotropic material Hui and Ruina (1995) show analytically that the A_n coefficients for $n < -1$ are nonzero. However, in the limit as the inelastic zone shrinks to zero these coefficients go to zero. Thus eliminating terms in the series



Stress Intensity Factors, Fig. 1 Crack tip coordinate system

solution that are more singular than $z^{-1/2}$, noting that as $r \rightarrow 0$ the $z^{-1/2}$ term dominates, and using the definition of the Mode-III stress intensity factor, the crack tip fields for the Mode-III problem are given by

$$\begin{pmatrix} \sigma_{13} \\ \sigma_{23} \end{pmatrix} = \frac{K_{III}}{\sqrt{2\pi r}} \begin{pmatrix} -\sin \frac{\theta}{2} \\ \cos \frac{\theta}{2} \end{pmatrix} \text{ as } r \rightarrow 0, \quad (2)$$

and

$$u_3 = \sqrt{\frac{2r}{\pi}} \frac{K_{III}}{\mu} \sin \frac{\theta}{2}. \quad (3)$$

The stress intensity factor, K_{III} is not determined from this analysis. In general K_{III} will depend linearly on the applied loads and will also depend on the specific geometry of the cracked body and on the distribution of loads.

Similar analyses can be performed for the plane-strain or plane-stress Mode-I and Mode-II problems (Rice 1968). The Mode-I stress fields are

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{12} \\ \sigma_{22} \end{pmatrix} = \frac{K_I}{\sqrt{2\pi r}} \cos \frac{\theta}{2} \begin{pmatrix} 1 - \sin \frac{\theta}{2} \sin \frac{3\theta}{2} \\ \sin \frac{\theta}{2} \cos \frac{3\theta}{2} \\ 1 + \sin \frac{\theta}{2} \sin \frac{3\theta}{2} \end{pmatrix}, \quad (4)$$

with displacements

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \frac{K_I}{2\mu} \sqrt{\frac{r}{2\pi}} \begin{pmatrix} \cos \frac{\theta}{2} (\kappa - \cos \theta) \\ \sin \frac{\theta}{2} (\kappa - \cos \theta) \end{pmatrix}, \quad (5)$$

where $\kappa = 3 - 4\nu$ for plane strain and $\kappa = (3 - \nu)/(1 + \nu)$ for plane stress.

The Mode-II fields are

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{12} \\ \sigma_{22} \end{pmatrix} = \frac{K_{II}}{\sqrt{2\pi r}} \begin{pmatrix} -\sin \frac{\theta}{2} (2 + \cos \frac{\theta}{2} \cos \frac{3\theta}{2}) \\ \cos \frac{\theta}{2} (1 - \sin \frac{\theta}{2} \sin \frac{3\theta}{2}) \\ \sin \frac{\theta}{2} \cos \frac{\theta}{2} \cos \frac{3\theta}{2} \end{pmatrix}, \quad (6)$$

with displacements

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \frac{K_{II}}{2\mu} \sqrt{\frac{r}{2\pi}} \begin{pmatrix} \sin \frac{\theta}{2} (\kappa + 2 + \cos \theta) \\ -\cos \frac{\theta}{2} (\kappa - 2 + \cos \theta) \end{pmatrix}. \quad (7)$$

The material will yield or otherwise deform nonlinearly to eliminate the predicted infinite stress, thus very near the crack tip the above expressions are not

accurate descriptions of the stress field. However, if r_p the size of the zone near the crack tip in which inelastic deformation occurs is small relative to the crack length, a , the stress outside of this “yielding zone” will be well approximated by the above. This is the so-called “small scale yielding” (SSY) assumption in fracture (Rice 1968). In the context of SSY all of the external loading and problem geometry are reflected in the values of the stress intensity factors. The crack tip stresses are uniquely determined by the stress intensity factors.

Key Applications

The key applications of stress intensity factors are for the calculation of the energy flows in fracture processes and in the prediction of crack initiation, crack growth rate, and crack path evolution.

Energy Release Rate

When a crack grows, the stresses that were acting across the newly created fracture surfaces are released and the new fracture surfaces move relative to each other. During this process negative work is done, or put another way, the body releases elastic strain energy as the crack grows. The energy released per unit of new fracture surface is called the “energy release rate,” labeled as G with units of $[F \cdot L / L^2]$. For a crack of arbitrary geometry and crack growth onto an arbitrary new surface, G can be difficult to calculate, however, for the simple case in which a two-dimensional crack grows straight ahead, i.e., in the x_1 direction with respect to Fig. 1, the energy release rate is given by (Irwin 1957)

$$G = \lim_{\Delta a \rightarrow 0} \frac{1}{\Delta a} \int_0^{\Delta a} \sigma_{i2}(x_1, 0) u_i (\Delta a - x_1, \pi) dx_1, \quad (8)$$

where Δa is the increment of crack length and the arguments of σ and u are in polar coordinates. Note that $u_1(\Delta a - x_1, \pi)$ represents the crack opening displacement under Mode-I loading while u_2 and u_3 represent the crack sliding displacements under Mode-II and Mode-III loadings, respectively. Using the stress and displacement fields given above, G is related to the stress intensity factors by

$$G = \frac{K_I^2}{E'} + \frac{K_{II}^2}{E'} + \frac{K_{III}^2}{2\mu}, \quad (9)$$

where $E' = E$ for plane stress and $E' = E/(1 - \nu^2)$ for plane strain.

Equating the work done on a body to the energy released per unit new fracture surface area it can be shown that

$$G = \frac{P^2}{2b} \frac{\partial c}{\partial a}, \quad (10)$$

where, P is the applied load, c is the compliance, and, in a two-dimensional problem, b is the plate thickness. This relationship can be used to determine G (and hence stress intensity factors) if the compliance can be estimated from structural analysis or measurements.

Calculation of Stress Intensity Factors

Stress intensity factors are calculated through the use of elasticity theory, handbooks of tabulated solutions, compliance methods, and computational methods.

In relatively few practical problems can the stress intensity factor be computed in closed form. However, one example is a straight, finite crack small enough to be considered as embedded in an infinite body. If the crack is loaded on its surfaces by a distribution of tractions $t_i = P_i(x_1)$ on the top surface and equal and opposite tractions $t_i = -P_i(x_1)$ on the bottom surface, see Fig. 2, then the stress intensity factors for the right-hand crack tip can be calculated as

$$\begin{Bmatrix} K_I \\ K_{II} \\ K_{III} \end{Bmatrix} = \frac{1}{\sqrt{\pi a}} \int_{-a}^a \left[\frac{a+t}{a-t} \right]^{1/2} \begin{Bmatrix} p_2(t) \\ p_1(t) \\ p_3(t) \end{Bmatrix} dt. \quad (11)$$

In the case that the tractions are constant the stress intensity factors will be

$$K_I = p_2 \sqrt{\pi a} \quad (12)$$

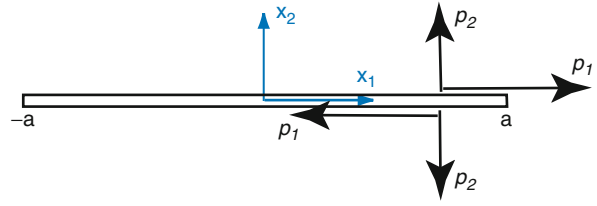
$$K_{II} = p_1 \sqrt{\pi a} \quad (13)$$

$$K_{III} = p_3 \sqrt{\pi a}. \quad (14)$$

If the crack surfaces are stress free, but the cracks are embedded in a stress field whose components are approximately constant away from the crack, then the stress intensity factors are the same as the above, replacing the p_i by the far-field stresses, $p_2 = \sigma_{22}$, $p_1 = \sigma_{12}$, $p_3 = \sigma_{23}$.

Stress intensity factors have been calculated for a great many problems involving both laboratory test specimens and practical problems in which one wishes to assess the fracture safety of a structure or component. These results are tabulated in handbooks of stress intensity factors (Tada et al. 2000), (Murakami and Aoki 1987). A selection of stress intensity factor solutions is given in Table 1.

The most flexible approach to the computation of K_I and G is the finite element method. In most cases the crack is modeled explicitly. Once the stress analysis is



Stress Intensity Factors, Fig. 2 Definition of tractions acting on the surface of a finite crack in an infinite body

performed, K_I or G can be determined through a number of post-processing steps. The conceptually simplest approach is to fit the stress or displacement fields near the crack tip to (2–7) Chan et al. (1970). The crack tip energy release rate can be calculated using the method of modified crack closure in which the work released during crack growth is calculated from the product of the reaction forces at the crack tip nodes and nodal displacements just behind the crack tip (Rybicki and Kanninen 1977). Separating the contributions to the energy due to Mode-I, -II, -III fracture and using the relation between energy release rate and stress intensity factors, the stress intensity factors can be calculated.

Stress intensity factors can also be calculated by the compliance method. Consider the double-cantilevered beam geometry shown in Fig. 3. The cracked portion is treated as two cantilevered beams of height h and thickness b loaded with force P . From beam theory the deflection at the end of a cantilevered beam of length a is $u = Pa^3/3EI$, where $I = bh^3/12$. Thus the compliance is $c(a) = 2 \cdot u/P = 8a^3/Ebh^3$ (since there are two beams). From (10), $G = \frac{P^2}{2b} \frac{\partial c}{\partial a} = 12 \left(\frac{P}{b} \right)^2 \frac{a^2}{Eh^3}$.

Fracture Criteria Based on Stress Intensity Factors

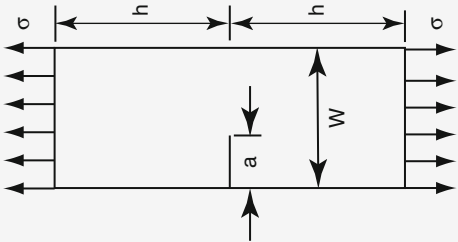
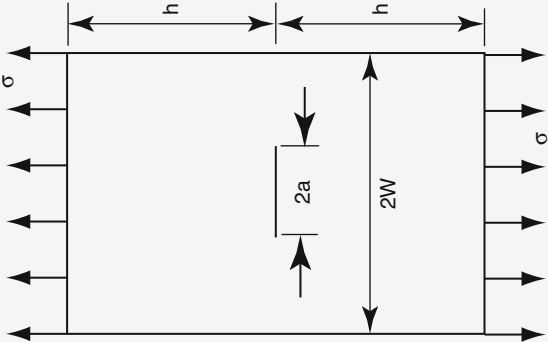
For a stationary, sharp crack under mode-I loading, the crack tip deformation and failure is driven solely by K_I . Thus it can be postulated that the crack will grow when

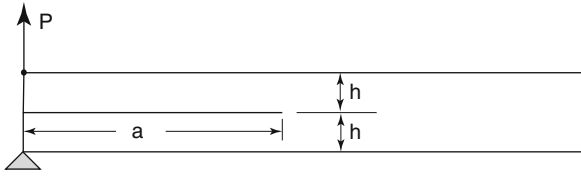
$$K_I \geq K_C, \quad (15)$$

where K_I is the stress intensity factor due to the applied load and K_C is a critical value that represents the resistance of the material to fracture. The value of K_C can be considered a material property in that, although it may depend on thickness and environmental factors, it can be measured and tabulated (DoD 1998).

Under mixed-mode loading the crack initiation criteria may be based on a surface of K_I , K_{II} , and K_{III}

Stress Intensity Factors, Table 1 Stress intensity solutions for several fracture test specimen geometries. $E' = E$ (plane stress), $E' = E/(1 - \nu^2)$ (plane strain) (Adapted from (Tada et al. 2000) and (ASTM 2005))

	<p>Single Edge Notch Tension (SENT)</p> <p>$h/W > 1$</p> $K_I = \sigma \sqrt{\pi a} F(a/W)$ $F(a/W) = 0.265(1 - a/W)^4 + \frac{.857 + .265a/W}{(1 - a/W)^{3/2}}$
	<p>Center Cracked Tension (CCT)</p> <p>$h/W > 3$</p> $K_I = \sigma \sqrt{\pi a} F(a/W)$ $F(a/W) = \sqrt{\sec \frac{\pi a}{2W} [1 - 0.025(a/W)^2 + .06(a/W)^4]}$



Stress Intensity Factors, Fig. 3 Double cantilever beam geometry

values. In isotropic materials cracks under mixed-mode loading will generally evolve to surfaces for which the loading is pure Mode-I.

Cross-References

- [Crack Growth in Brittle and Ductile Solids](#)
- [Crack Growth in Noncrystalline Solids](#)
- [Crack Initiation in Brittle Solids](#)
- [Griffith Theory of Fracture](#)
- [Modes of Fracture](#)

References

- ASTM, ASTM E 1820: Standard Test Method for Measurement of Fracture Toughness. ASTM International, West Conshohocken, (2005)
- S. Chan, I. Tuba, W. Wilson, On the finite element method in linear fracture mechanics. *Eng. Fract. Mech.* **2**, 1–17 (1970)
- DoD, Metallic materials and elements for aerospace vehicle structures. U.S. Dept. of Defense (1998)
- C.Y. Hui, A. Ruina, Why K? High order singularities and small scale yielding. *Int. J. Fract.* **72**, 97–120 (1995)
- G. Irwin, Analysis of stresses and strains near the end of a crack traversing a plate. *J. Appl. Mech.* **24**, 361–364 (1957)
- Y. Murakami, S. Aoki, *Stress Intensity Factors Handbook* (Pergamon, Oxford, NY, 1987)
- J.R. Rice, Mathematical Analysis in the Mechanics of Fracture, in *Fracture*, chapter 3, ed. by H. Liebowitz, vol. II (Academic, New York, 1968), pp. 191–311
- E. Rybicki, M. Kanninen, A finite element calculation of stress intensity factors by a modified crack closure integral. *Eng. Fract. Mech.* **9**, 931–938 (1977)
- H. Tada, P.C. Paris, G.R. Irwin, *The Stress Analysis of Cracks Handbook* (ASME Press, New York, 2000)

Stress Relaxation in Bolted Joints

- [Fastener Failure and Safety Related Issues](#)

Stresses in Contacting Materials

W. WAYNE CHEN

Mechanical Engineering, Northwestern University,
Evanston, IL, USA

Synonyms

[Contact stress analysis](#)

Definition

Contact stresses are the elastic stresses in a half-space produced by the surface normal and tangential tractions as a result of the interfacial contact.

Scientific Fundamentals

Stresses in a Cylindrical Contact (Line Contact)

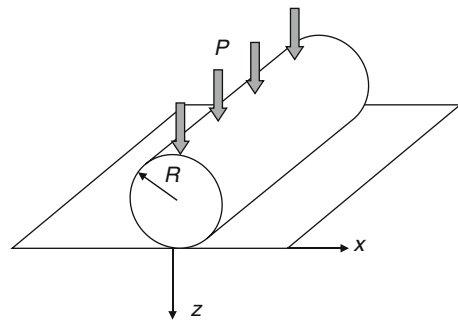
Figure 1 shows the contact of a cylinder and a half-space (or two cylinders), where a line distributed load, P , is applied along the entire infinite axis of the cylinder. The ideal contact area is a narrow strip and the strip width in the x - z plane is $2a_0$. The resulting contact pressure distribution is a parabolic function in terms of the x coordinate.

$$p(x) = p_0 \sqrt{1 - x^2/a_0^2}$$

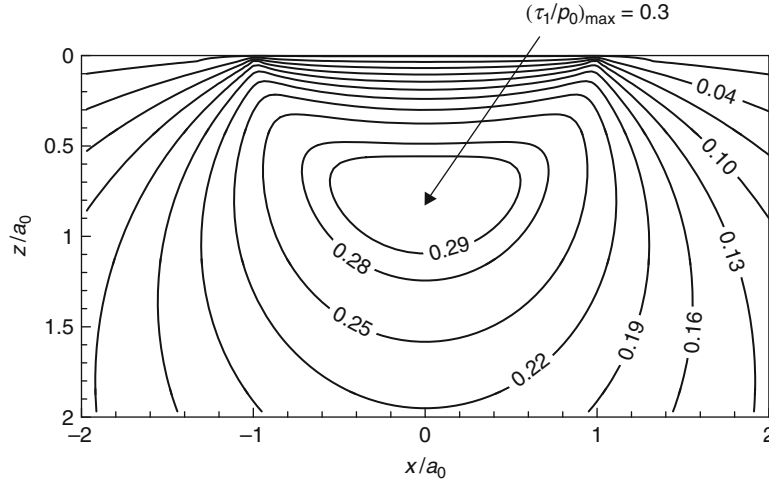
$$p_0 = 2P/\pi a_0 \quad (1)$$

where p_0 is the peak contact pressure.

For a plane strain state, the stresses in each cross section have the same distributions. The stresses in a subsurface point (x, z) can be expressed by (Johnson 1985),



Stresses in Contacting Materials, Fig. 1 A cylinder in a contact with a half-space



Stresses in Contacting Materials, Fig. 2 Contour of the principal shear stress, τ_1/p_0 , caused by a line contact in the cross section

$$\sigma_{xx} = -\frac{p_0}{a_0} \left[m \left(1 + \frac{z^2 + n^2}{n^2 + m^2} \right) - 2z \right]$$

$$\sigma_{zz} = -\frac{p_0}{a_0} m \left(1 - \frac{z^2 + n^2}{n^2 + m^2} \right), \tau_{xz} = -\frac{p_0}{a_0} n \left(\frac{m^2 - z^2}{n^2 + m^2} \right). \quad (2)$$

where

$$m^2 = \frac{1}{2} \left[\sqrt{(a_0^2 - x^2 + z^2)^2 + 4x^2 z^2} + (a_0^2 - x^2 + z^2) \right]$$

$$n^2 = \frac{1}{2} \left[\sqrt{(a_0^2 - x^2 + z^2)^2 + 4x^2 z^2} - (a_0^2 - x^2 + z^2) \right].$$

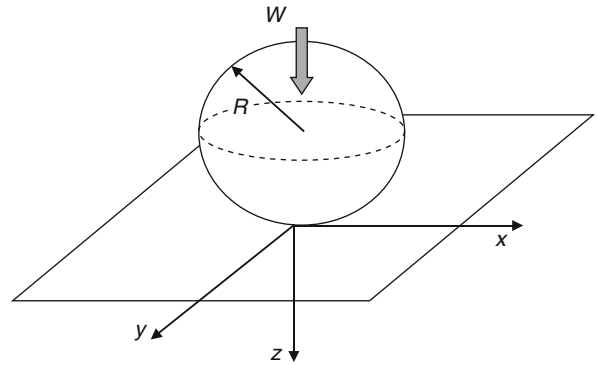
These stresses are independent of Poisson's ratio of the investigated body, and the normal stress along the cylinder axis is $\sigma_{yy} = \nu(\sigma_{xx} + \sigma_{zz})$. Figure 2 shows the contour of principal shear stress, τ_1 (see the definition in (3)), in the cross section perpendicular to the cylinder axis.

$$\tau_1 = \frac{1}{2} \sqrt{(\sigma_{xx} - \sigma_{zz})^2 + 4\tau_{xz}^2} \quad (3)$$

The maximum principal shear stress is $(\tau_1/p_0)_{\max} = 0.3$ and locates at the depth of $z = 0.78a_0$.

Stresses in a Spherical Contact (Point Contact)

A point contact between a sphere and a flat is given in Fig. 3, where the contact dimensions are much smaller than the contacting bodies' geometries. Normal load, W , is applied along the z axis and the origin is the initial



Stresses in Contacting Materials, Fig. 3 A sphere in a contact with a half-space

contact point. The interfacial contact occurs in a circle area with a radius of a_0 , and the contact pressure, p , is,

$$p(x, y) = p_0 \sqrt{1 - x^2/a_0^2 - y^2/a_0^2}$$

$$p_0 = 3W/2\pi a_0^2 \quad (4)$$

where p_0 is the peak contact pressure. If a tangential force pushes the sphere to sliding along the x axis, the shear traction may occur at the contact region. Following the Coulomb friction law, the shear traction, q_x , can be expressed by,

$$q_x(x, y) = \mu_f p_0 \sqrt{1 - x^2/a_0^2 - y^2/a_0^2} \quad (5)$$

Here, μ_f is the friction coefficient.

The stress field in a half-space produced by the normal and tangential tractions in (4) and (5) was derived by Hamilton (1983). With the normal pressure alone, the stresses along the z axis ($x = y = 0$) are,

$$\begin{aligned}\sigma_{xx} = \sigma_{yy} &= \frac{p_0}{a_0} \left[(1 + \nu) \left(z \arctan\left(\frac{a_0}{z}\right) - a_0 \right) + \frac{a_0^3}{2(a_0^2 + z^2)} \right] \\ \sigma_{zz} &= -\frac{p_0}{a_0} \frac{a_0^3}{(a_0^2 + z^2)}, \sigma_{xy} = \sigma_{xz} = \sigma_{yz} = 0\end{aligned}\quad (6)$$

The stresses at the surface ($z = 0$) are,

$$\begin{aligned}\sigma_{xx} &= \begin{cases} \frac{p_0}{a_0} \left[\frac{(y^2 - x^2)(1 - 2\nu)}{3r^4} \{ (a_0^2 - r^2)^{3/2} - a_0^3 \} - \frac{(x^2 + 2\nu y^2)\sqrt{a_0^2 - r^2}}{r^2} \right] & r \leq a_0 \\ \frac{p_0}{a_0} \left[\frac{(x^2 - y^2)(1 - 2\nu)a_0^3}{3r^4} \right] & r > a_0 \end{cases} \\ \sigma_{yy} &= \begin{cases} \frac{p_0}{a_0} \left[\frac{(x^2 - y^2)(1 - 2\nu)}{3r^4} \{ (a_0^2 - r^2)^{3/2} - a_0^3 \} - \frac{(y^2 + 2\nu x^2)\sqrt{a_0^2 - r^2}}{r^2} \right] & r \leq a_0 \\ \frac{p_0}{a_0} \left[\frac{(y^2 - x^2)(1 - 2\nu)a_0^3}{3r^4} \right] & r > a_0 \end{cases} \\ \sigma_{zz} &= \begin{cases} -\frac{p_0}{a_0} \sqrt{a_0^2 - r^2} & r \leq a_0 \\ 0 & r > a_0 \end{cases} \\ \sigma_{xy} &= \begin{cases} \frac{p_0}{a_0} \left[\frac{xy(1 - 2\nu)}{r^4} \{ 2a_0^3/3 - (a_0^2 - r^2)^{1/2} [r^2 + 2(a_0^2 - r^2)/3] \} \right] & r \leq a_0 \\ \frac{p_0}{a_0} \left[\frac{2xy(1 - 2\nu)a_0^3}{3r^4} \right] & r > a_0 \end{cases} \\ \sigma_{xz} &= \sigma_{yz} = 0\end{aligned}\quad (7)$$

Here, $r^2 = x^2 + y^2$. The stresses only depend on Poisson ratio of material if the pressure distribution is known. For the general solutions of stresses at any point (x, y, z) in the contacting half-space caused by the Hertz normal and tangential tractions, readers may refer to the literature (Hamilton 1983).

Contours of the von Mises equivalent stress σ_{VM} (see definition in (8)) and the first principal stress σ_1 ($\sigma_1 \geq \sigma_2 \geq \sigma_3$ are three principal stresses) under frictionless and frictional ($\mu_f = 0.3$) are shown in Fig. 4, where Poisson ratio is $\nu = 0.3$.

$$\sigma_{VM} = \sqrt{3S_{ij} : S_{ij}/2} \quad (8)$$

where $S_{ij} = \sigma_{ij} - \sigma_{kk}\delta_{ij}/3$ is the deviatoric stress. For a frictionless contact, the maximum von Mises stress is about $0.62p_0$ and occurs at the depth of $z = 0.48a_0$, which means that the onset of plastic deformation happens

below the surface. The tensile first principal stress σ_1 concentrates in a shallow surface region (with the maximum value at the contact circle edge), which is responsible for the surface ring (cone) cracks of brittle materials under the indentation load. When the sliding friction is considered ($\mu_f = 0.3$), the von Mises stress increases to $0.66p_0$ and moves closer to the surface ($z = 0.42a_0$). The magnitude and region of tensile first principal stress increases significantly. The trailing edge of sliding indenter is vulnerable to the transverse cracks.

The variations of local maximum values of the von Mises stress in the subsurface region and those at the contact surface with the friction coefficient were discussed by Hills et al. (1993), and are presented in Fig. 5. The maximum von Mises stress moves up to the surface when the friction coefficient reaches about 0.3.

Stresses in a Rigid Flat Punch Indentation

Figure 6 shows a frictionless rigid flat punch is in contact with a half-space. The punch has a cylindrical shape with a radius a_0 , thus a cylindrical coordinate is used for this axisymmetric problem. The pressure on the contact end of punch is (Johnson 1985),

$$p(r) = \frac{p_m}{2\sqrt{1 - r^2/a_0^2}} \quad (9)$$

where $p_m = W/\pi a_0^2$ is the average contact pressure and W is the applied normal load.

The stresses in the semi-infinite body can be expressed by Sneddon (1946),

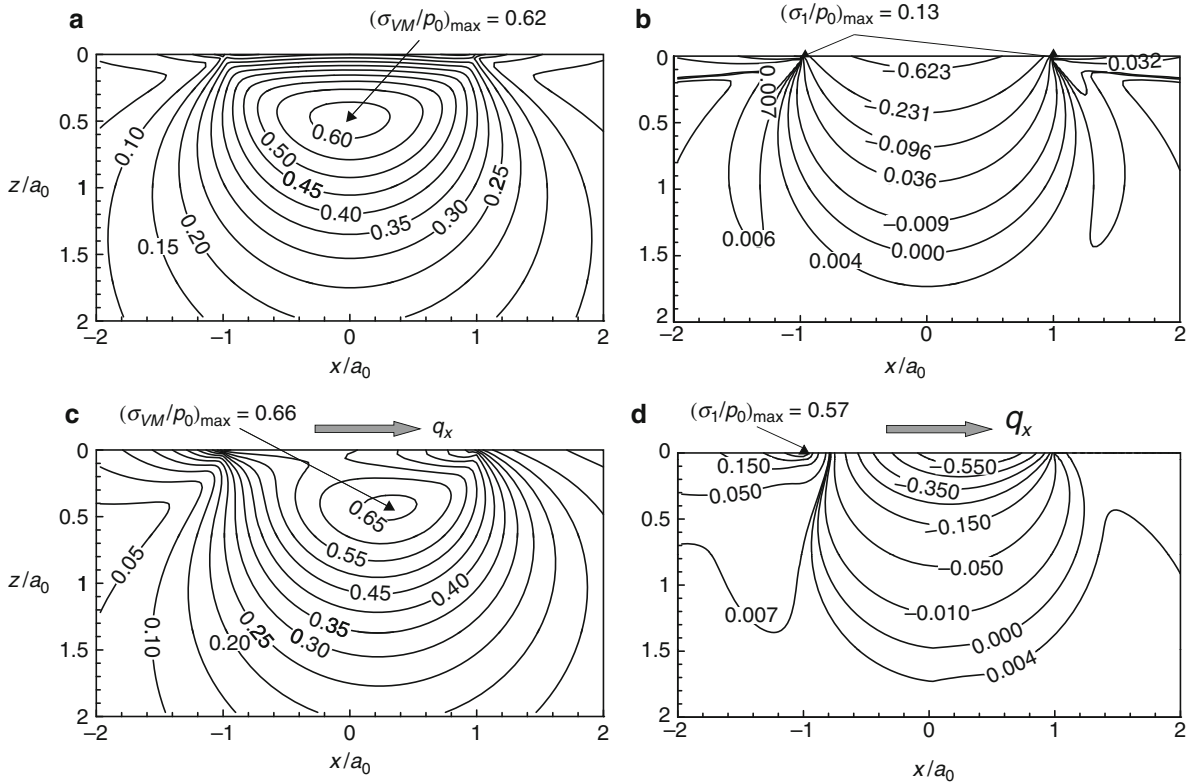
$$\begin{aligned}\sigma_{zz} &= -\frac{p_m}{2} (J_1^0 + \xi J_2^0), \sigma_{zr} = -\frac{p_m}{2} \xi J_2^1, \sigma_{z\theta} = \sigma_{r\theta} = 0 \\ \sigma_{\theta\theta} &= -p_m \nu J_0^1 - \frac{p_m}{\gamma} \left(\frac{1 - 2\nu}{2} J_0^1 - \xi J_2^1 \right), \\ \sigma_{rr} + \sigma_{\theta\theta} &= -\frac{p_m}{2} [(1 + 2\nu)J_1^0 - \xi J_2^0]\end{aligned}\quad (10)$$

where J_n^m denotes the integrals

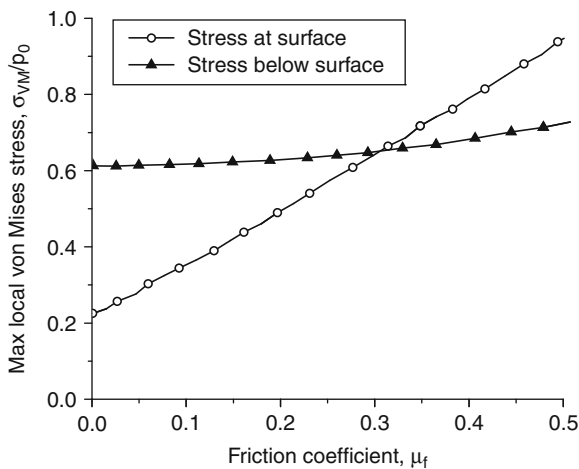
$$J_n^m = \text{Im} \left[\int_0^\infty \eta^{n-1} \exp\{-\eta(\xi - i)\} J_m(\gamma\eta) d\eta \right] \quad (11)$$

Here, J_m is the Bessel function of the first kind, and $\gamma = r/a_0$ and $\xi = z/a_0$. The integrals J_n^m can be evaluated using the associated Legendre functions (Sneddon 1946).

From (9), the pressure reaches an infinite value at the edge of the contact area ($r = a_0$). Figure 7 shows the von Mises stress contour in the cross section ($\theta = 0$). A singular value of von Mises stress can be found near the punch edge as well. Thus, for an indentation with an ideal flat-end punch, the material plasticity may occur even at the lightest load.



Stresses in Contacting Materials, Fig. 4 Stress contours in the cross section of $y = 0$, (a) von Mises stress, σ_{VM}/p_0 , at $\mu_f = 0$, (b) first principal stress, σ_1/p_0 , at $\mu_f = 0$, (c) von Mises stress, σ_{VM}/p_0 , at $\mu_f = 0.3$, and (d) first principal stress, σ_1/p_0 , at $\mu_f = 0.3$



Stresses in Contacting Materials, Fig. 5 Maximum von Mises stresses at surface and below surface as a function of friction coefficient (Hills et al. 1993)

Stresses by Distributed Surface Traction

Contact pressure and shear tractions may be in any arbitrary distribution other than those aforementioned. For a problem with distributed normal contact pressure, the stress solution of a unit concentrated normal load applied at the origin (see Fig. 8a) can be treated as the fundamental solution. The stress components at any point (x, y, z) in the half-space are expressed by the well-known Boussinesq solutions (Johnson 1985).

$$T_{xx}^N(x, y, z) = \frac{1}{2\pi} \left[\frac{1-2\nu}{r^2} \left\{ \left(1 - \frac{z}{\rho} \right) \frac{x^2 - y^2}{r^2} + \frac{zy^2}{\rho^3} \right\} - \frac{3zx^2}{\rho^5} \right]$$

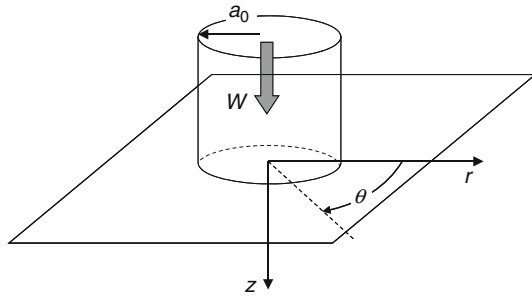
$$T_{yy}^N(x, y, z) = T_{11}^N(y, x, z)$$

$$T_{zz}^N(x, y, z) = -\frac{3}{2\pi} \frac{z^3}{\rho^5}$$

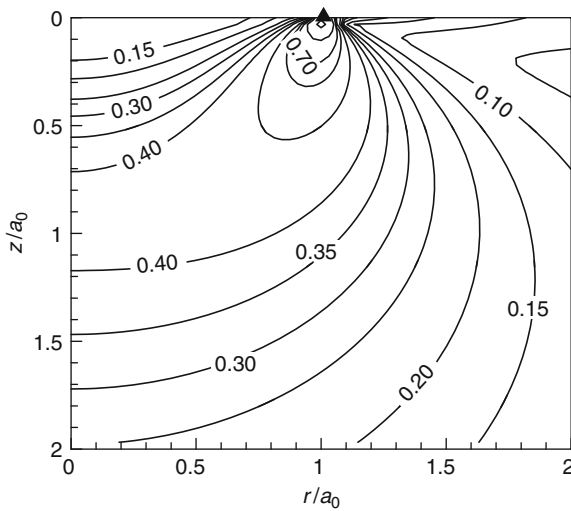
$$T_{xy}^N(x, y, z) = \frac{1}{2\pi} \left[\frac{1-2\nu}{r^2} \left\{ \left(1 - \frac{z}{\rho} \right) \frac{xy}{r^2} - \frac{xyz}{\rho^3} \right\} - \frac{3xyz}{\rho^5} \right]$$

$$T_{xz}^N(x, y, z) = T_{yz}^N(y, x, z) = -\frac{3}{2\pi} \frac{xz^2}{\rho^5} \quad (12)$$

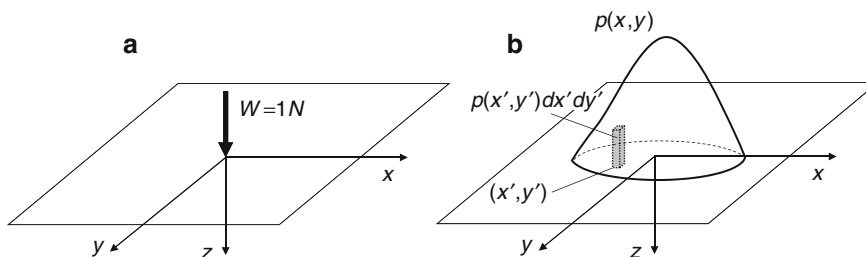
where $r^2 = x^2 + y^2$ and $\rho = \sqrt{x^2 + y^2 + z^2}$. According to the superposition principle of the linear system,



Stresses in Contacting Materials, Fig. 6 A rigid flat punch in a frictionless contact with a half-space



Stresses in Contacting Materials, Fig. 7 Contour of von Mises stress, σ_{VM}/p_m , in the plane of $\theta = 0$ ($\nu = 0.3$)



Stresses in Contacting Materials, Fig. 8 (a) A concentrated normal load on a half-space, and (b) a distributed pressure applied on a half-space

the stresses due to a distributed pressure $p(x, y)$ can be written as,

$$\sigma_{ij}(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T_{ij}^N(x-x', y-y', z) \cdot p(x', y') dx' dy' \quad (13)$$

Equation 13 can be evaluated through a numerical approach, where the half-space is subdivided into small elements. Assume contact pressure is uniform in each surface element. Therefore, a real pressure distribution is described by a piecewise constant function. Stress components caused by the unit uniform pressure in one element are first calculated and kept as the influence coefficients (IC), which can be used in the future stress computation in conjunction with the sampled pressure vector. The details on mesh and influence coefficients were discussed by Liu and Wang (2002). Fundamental solutions of stresses due to the shear traction were also given in the Johnson's book (Johnson 1985). An example of distributed pressure as a result of a rough surface contact (in Fig. 9a) was given by Liu and Wang (2002). The contour of the von Mises stress is also shown in Fig. 9b.

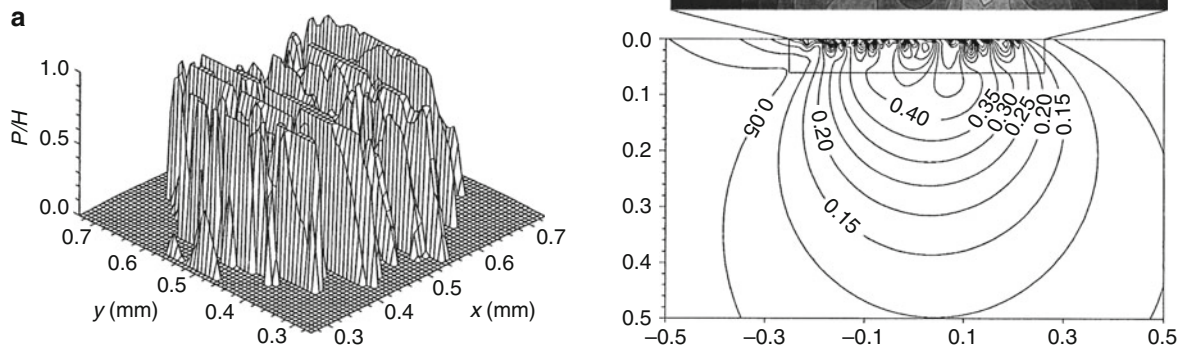
Key Applications

Analyzing stresses in contacting materials is of significant importance to the design of contacting components. The knowledge of stress field provides a basis for the investigation of surface-related failures, such as crack, wear, material yielding, and fatigue.

The analytical solutions of stresses in a point contact can be used to determine the critical load leading to the onset of plasticity under the spherical indenter (Jackson and Green, 2005). Based on (6), the equivalent von Mises stress along the depth can be written as,

$$\frac{\sigma_{VM}}{p_0} = \frac{3}{2(1+z^2/a_0^2)} - (1+\nu) \left[1 - \frac{z}{a_0} \arctan\left(\frac{a_0}{z}\right) \right] \quad (14)$$

For the material with Poisson ratio, $\nu = 0.3$, the critical normal load was solved as (Jackson and Green 2005),



Stresses in Contacting Materials, Fig. 9 (a) Pressure distribution caused by a contact with rough surface (RMS roughness, $R_q = 0.21 \mu\text{m}$ and material hardness, $H = 1.8 \text{ GPa}$), and (b) contour of the von Mises stress, σ_{VM}/H , in the half-space (friction coefficient, $\mu_f = 0.25$) (Liu and Wang, 2002)

$$W_c = \frac{(1.6Y\pi)^3 R^2}{6E^*2} \quad (15)$$

where Y is the material yielding point.

The stress values in contacting materials were also utilized to study the fatigue failure in the mixed lubrication contact (Epstein et al. 2003). Based on the survival probability model, the number of load cycles leading to fatigue, N , is inversely proportional to the integration of equivalent stress (the von Mises stress was generally used) over the entire volume.

$$\ln \frac{1}{S} \sim N^e \iiint_{\Omega} (\sigma_{VM})^{e-c} d\Omega \quad (16)$$

Here, S is the survival probability after N th cycle, and e and c are material constants.

Cross-References

- [Contact of Layered Materials](#)
- [Influence Coefficients for Contact Mechanics](#)
- [Punch Contact Theories Including the Edge Effect](#)

References

- D. Epstein et al., Effect of surface topography on contact fatigue in mixed lubrication. *Tribol. Trans.* **46**, 506 (2003)
- G.M. Hamilton, Explicit equations for the stresses beneath a sliding spherical contact. *Proc. Inst. Mech. Eng.* **197**, 53 (1983)
- D.A. Hills et al., *Mechanics of Elastic Contact* (Butterworth-Heinemann, Oxford, 1993)
- R.L. Jackson, I. Green, A finite element study of elasto-plastic hemispherical contact against a rigid flat. *ASME J. Tribol.* **127**, 343 (2005)

K.L. Johnson, *Contact Mechanics* (Cambridge University Press, London, 1985)

S.B. Liu, Q. Wang, Study contact stress fields caused by surface tractions with a discrete convolution and fast Fourier transform algorithm. *ASME J. Tribol.* **124**, 36 (2002)

I.N. Sneddon, Boussinesq's problem for a flat-ended cylinder. *Proc. Camb. Philos. Soc.* **42**, 29 (1946)

Stress-Induced Lubricant Degradation and Viscosity Loss

ILYA I. KUDISH

Department of Mathematics, Kettering University, Flint, MI, USA

Synonyms

[Degradation of polymeric viscosity improvers](#); [Lubricant viscosity loss due to additive degradation](#); [Stress-induced polymeric additive degradation](#)

Definition

All modern fluid lubricants are formulated, that is, they are represented by a base oil with added various additives that improve some desirable lubricant properties and retard some negative effects. One of the most widely used classes of additives is viscosity improvers/modifiers. Their main purpose is to reduce lubricant viscosity at low temperatures and to prevent lubricant viscosity from significant reduction at high temperatures

(i.e., to maintain lubricant viscosity at the level sufficient for normal operation of machines and mechanisms). However, due to stresses experienced by lubricants these polymer additives gradually degrade (polymer chains break and the polymer molecular distribution changes), which changes lubricant properties such as its viscosity and leads to negative effects in lubricated mechanisms. Stress-induced degradation of polymer additives depends on the polymer properties, structure and molecular weight, the motion lubricant is involved in, and the stresses it experiences.

Scientific Fundamentals

In lubricated contacts, oil film thickness is a critical design feature of a lubricant and it is controlled, in large part, by viscosity. Should viscosity drop below a critical value, opposite surfaces can begin to come into contact, resulting in premature wear and fatigue damage. High-molecular-weight polymers, known as viscosity modifiers (VM) or viscosity improvers (VI), are added to lubricating oils to boost viscosity at high temperatures while minimizing thickening contribution at low temperatures. Viscosity increase is proportional to both VM concentration and molecular weight. To preserve lubricant film thickness between the moving surfaces over time, it is desirable to minimize molecular weight degradation of the polymer additive.

There exists a large body of experimental and theoretical studies of polymer degradation kinetics. Most of the theoretical studies of changes in the length distribution of polymer molecules over time are concerned with polymer degradation caused by radiation (Saito 1958) and thermal degradation of a polymer dissolved in a fluid (Montroll and Simha 1940; Ziff and McGrady 1985, 1986). Usually, in studies of thermal degradation, the mechanical stresses acting upon fluid were not considered. Qualitative theoretical and experimental studies of stress-induced lubricant degradation were conducted in (Odell et al. 1988; Covitch 1998). Here, a kinetics approach to stress-induced degradation of polymeric additives based on polymeric molecules with linear structure and star polymer molecules and their effect on lubricated contact parameters and its fatigue life are considered.

Modeling of Lubricant Degradation and Viscosity Loss

Consider polymer molecules with linear structure. Let $W(t, \mathbf{x}, l)$ be the distribution density of polymer molecule chains of length l (l is defined as the number of monomer units in the polymer chain, also known as the

degree of polymerization) at time moment t in a unit volume centered at \mathbf{x} . It is determined in such a way that $W(t, \mathbf{x}, l) \Delta l \Delta v$ is the weight of polymer molecules at time t located in a small fluid volume Δv centered at \mathbf{x} with molecule chain lengths from l to $l + \Delta l$. The weight of one polymer molecule of chain length l is $w = w_m l$, where w_m is the monomer molecular weight. Polymer molecules are considered to be stretched along the flow streamlines. A detailed derivation and theoretical and numerical analysis of the polymer degradation problem reduced to a kinetic equation

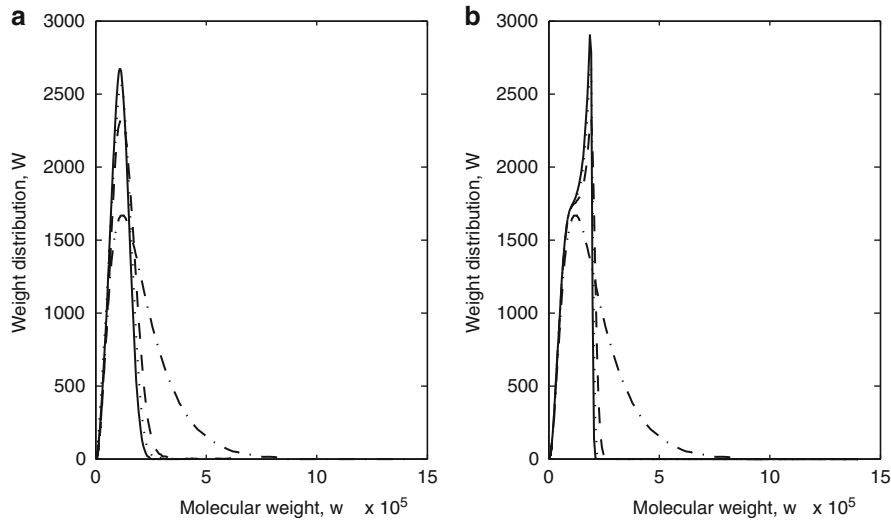
$$\rho \left(\frac{\partial}{\partial t} + \mathbf{u} \frac{\partial}{\partial \mathbf{x}} \right) \frac{W}{\rho} = \frac{2l}{\tau_f} \int_l^\infty R(t, L) p_c(t, l, L) W(t, \mathbf{x}, L) \frac{dL}{L} - \frac{1}{\tau_f} R(t, \mathbf{x}, l) W(t, \mathbf{x}, l), \quad W(0, \mathbf{x}, l) = W^0(\mathbf{x}, l), \quad (1)$$

is done in (Kudish et al. 2002, 2003). Here ρ is the lubricant density, $\mathbf{u} = (u_x, u_y, u_z)$ is the vector of the velocity of the fluid flow, $W^0(\mathbf{x}, l)$ is the initial distribution density of polymer molecule chains, τ_f is the characteristic time of one act of polymer fragmentation, $R(t, l)$ is the probability of dissociation of a polymer molecule of a chain length l at a time moment t , and $p_c(t, l, L)$ ($l \leq L$) is the conditional probability density such that $p_c(t, l, L) \Delta l$ is the probability of a polymer molecule of chain length L at the time moment t to break into two fragments of chain lengths λ and $L - \lambda$, $l \leq \lambda \leq l + \Delta l$. The specific expressions for R and p_c are derived in (Kudish et al. 2003) and have the form

$$R(t, l) = 0 \text{ if } l \leq L_*, \\ R(t, l) = 1 - \left(\frac{1}{L_*} \right)^{\frac{2\alpha U_A}{kT}} \exp \left[-\frac{\alpha U_A}{kT} \left(\frac{l^2}{L_*^2} - 1 \right) \right] \text{ if } l \geq L_*, \quad (2)$$

$$L_* = \sqrt{\frac{\mu_a}{\mu}} L_0, \quad L_0 = \sqrt{\frac{U_A}{Ca_* l_*^2 \mu_a S}}, \quad U_A = \frac{U}{N_A}, \quad (3) \\ p_c(t, l, L) = \ln 2 \frac{4|L - 2l|}{L^2} \exp \left[-\ln 2 \frac{4l(L - l)}{L^2} \right],$$

where μ and μ_a are the lubricant viscosity and ambient lubricant viscosity, respectively, S is the shear strain rate, U is a C-C bond dissociation energy per polymer mole, N_A is the Avogadro number, $N_A = 6.022 \cdot 10^{23} \text{ mole}^{-1}$, T is the lubricant temperature, k is Boltzmann's constant ($k = 1.38 \cdot 10^{-23} \text{ J/K}$), a_* and l_* are the polymer molecule bead radius and bond length, respectively, and C and α are dimensionless shield constants.



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 1 Polymer molecular weight distributions at different time moments during testing of the lubricant with OCP10 VM (a) and obtained from numerical modeling (b) for $U = 347$ kJ/mole, $C = 0.044$, and $\alpha = 0.008$: dashed-dotted curve – initial molecular weight distribution, dashed curve – after 30 cycles, dotted curve – after 100 cycles, solid curve – after 250 cycles

The lubricant viscosity μ depends on the distribution of the polymer additive that is expressed by the empirical Huggins and Mark-Houwink equations (Billmeyer 1966; Crespi et al. 1977):

$$\mu = \mu_a \frac{1 + c_p[\eta] + k_H(c_p[\eta])^2}{1 + c_p[\eta]_a + k_H(c_p[\eta]_a)^2}, \quad [\eta] = k' M_w^\beta,$$

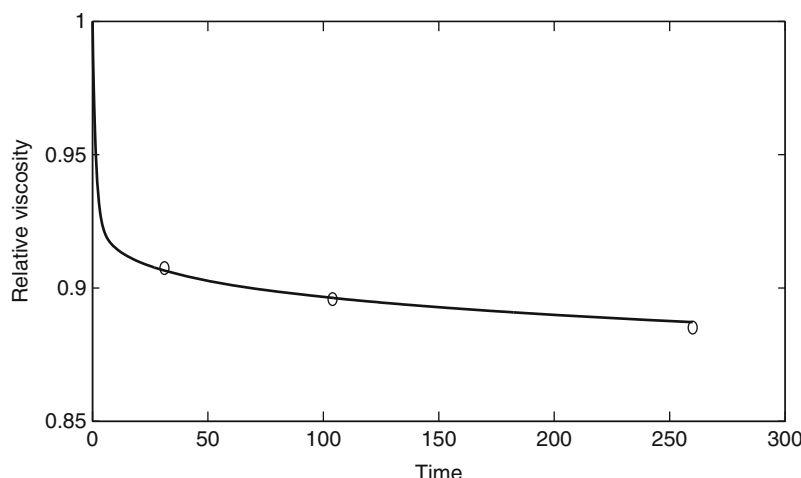
$$M_w = \left\{ \int_0^\infty w^\beta W(t, x, w) dw / \int_0^\infty W(t, x, w) dw \right\}^{1/\beta}, \quad (4)$$

where μ_a is the initial lubricant viscosity, c_p and $[\eta]$ are the polymer concentration and the intrinsic viscosity, respectively, $[\eta]_a$ is the intrinsic viscosity at the initial time moment, k_H is the Huggins constant, and k' and β are the Mark-Houwink constants.

Some basic properties of solutions of the kinetic equation (1)–(3) are established in (Kudishet al. 2002). In particular, it is shown that in an isolated system the polymer weight per unit mass of the lubricant is conserved in time along the lubricant flow streamlines while the number of polymer molecules per unit of the lubricant mass is a nondecreasing function of time. It is shown that the qualitative and quantitative behavior of numerically calculated polymer molecular weight distributions and viscosity loss are in excellent agreement with the corresponding test data. The examples of such a comparison for lubricant with OCP10 VM are

presented in Figs. 1 and 2 (Kudish et al. 2003) for the distribution of polymer molecular weight W and viscosity μ loss with time (number of cycles in Kurt Orbahn fuel injector bench test). The data in these figures is obtained for $U = 347$ kJ/mole, $l_* = 0.154$ nm, $a_* = 0.374$ nm, $T = 310$ K, $\mu_a = 0.00919$ Pa · s, $S = 5,000$ s⁻¹, $c_p = 0.86$ g/dL, $k_H = 0.2$, $k' = 2.7 \cdot 10^{-4}$ dL/(g(g/mole)^{-β}), and $\beta = 0.7$ (Saito 1958). The best match of numerical and test results is provided by $C = 0.044$ and $\alpha = 0.008$. The initial molecular weight distribution $W^0(l)$ is obtained from the test data (Covitch 1998).

The process of lubricant degradation can be understood based on the analysis of the characteristic polymer chain length L_* from (2) and the expression for the probability of scission $R(t, l)$ from (2). When the lubricant temperature T increases, the probability of polymer scission R is determined by the two competing processes: decrease of the slope of R and shift of R toward larger values of chain lengths l due to usual decrease of the lubricant viscosity μ with temperature T . Therefore, the probability of polymer scission R decreases with temperature T . The increase of the lubricant viscosity μ and μ_a (for example, due to increase of pressure p) and/or of the shear strain rate S cause L_* to decrease, which, in turn, accelerates scission. The physical explanation of this mechanism is clear because an increase in μ and/or S leads to a corresponding increase in stretching forces acting upon polymer molecules due to increased friction



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 2 Loss of the lubricant viscosity μ caused by polymer molecule degradation versus number of cycles for the lubricant with OCP10 VM, $U = 347$ kJ/mole, $C = 0.044$, and $\alpha = 0.008$. Circles indicate the experimentally measured relative viscosity of the lubricants with OCP10

between the polymer molecules and the surrounding lubricant. Finally, an increase in the value of the product $Ca_* l_*^2$ leads to the same effect as the increase in μ_a or S . An increase in the value of $\alpha U/(kT)$ leads to a steeper distribution of the probability of scission R . That results in faster scission of polymer chains.

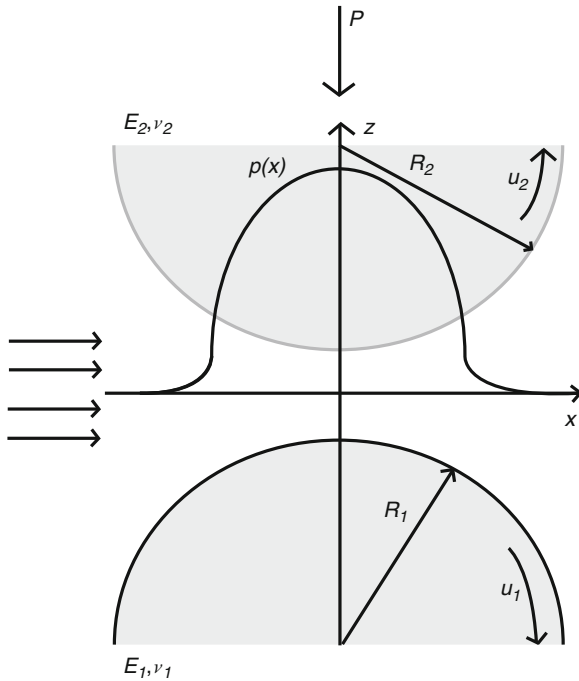
A model of stress-induced degradation of star polymers dissolved in lubricant is proposed and analyzed in (Kudish et al. 2005; Kudish 2007). Star polymer molecules are comprised of small organic cores to which a number of linear polymer arms are attached. The polymer degradation is modeled based on a system of kinetic equations for the distribution densities of star polymer molecules with different number of arms and arm chain lengths. Some properties of the solution are established. A numerical method for solution of the problem is proposed and realized. Some of the numerically simulated molecular weight distributions are compared with the independently obtained experimental ones. The lubricant viscosity losses due to polymer degradation are determined and compared with the experimentally measured ones. The theoretical and experimental data are in very good agreement.

It seems that star polymers of the same molecular weight as linear polymers are better shielded from stress-induced degradation than the linear ones (i.e., for a given molecular weight star polymer molecules undergo scission at a slower rate than polymer molecules with linear structure). This is due to the fact that in most cases the arms of a star polymer molecule are significantly shorter than the chain length of a linear polymer molecule of the same

molecular weight and, therefore, the probability R for them to break is lower. The latter means that the viscosity loss of a star polymer solution occurs more gradually than the viscosity loss of a linear polymer solution.

Elastohydrodynamic Lubrication by Formulated Lubricants that Undergo Stress-Induced Degradation

Modern lubricating oils are formulated with a variety of additives designed to (a) provide beneficial rheological characteristics to lubricants, (b) stabilize their physical and chemical properties, and (c) protect lubricated equipment against wear, fatigue, and corrosion. Under the influence of chemical and mechanical stresses and elevated temperatures lubricants tend to undergo certain reversible and irreversible changes. The reversible changes are caused by temporary alignment of polymeric additives in the direction of flow, resulting in an apparent drop in viscosity. When the liquid returns to a state of rest, the viscosity returns to its initial value. This is known as non-Newtonian rheology. The irreversible changes are due to a number of ongoing processes such as stress-induced scission of polymeric additives, oxidation, contamination, and so on. The latter detrimental processes limit the useful life of lubricants and can lead to costly repairs and down time if a lubricated system is not properly maintained. Here, the emphasis is placed on the combined effects of the lubricants' non-Newtonian rheology and stress-induced polymer molecule scission and on changes in lubricant contact parameters. Most lubricants exhibit some reversible non-Newtonian rheological properties (Bair and Winer 1979; Hoglund and



Stress-Induced Lubricant Degradation and Viscosity Loss,
Fig. 3 The general view of a lubricated contact

Jacobson 1986). Some rheological models of lubricant behavior are given in (Bair and Winer 1979; Eyring 1936).

Consider the application of the developed kinetics approach to the phenomenon of elastohydrodynamic lubrication (EHL) of surfaces lubricated with non-Newtonian fluids that undergo lubricant degradation. Consider a steady isothermal EHL problem for a line contact. Suppose two infinite parallel cylinders steadily move with linear surface speeds u_1 and u_2 . The cylinders have radii R_1 and R_2 and are made of elastic materials with Young's moduli E_1 and E_2 and Poisson's ratios ν_1 and ν_2 , respectively. The cylinders are separated by a thin lubrication layer and are loaded by the force P directed along their centers and normal to their axes (see Fig. 3).

The lubricant is assumed to be an incompressible fluid with a non-Newtonian rheology. The additive is assumed to be of linear structure. Under the classic EHL assumptions, such as (a) the motion is slow so that the inertia forces can be neglected in comparison with the viscous ones, (b) the lubrication film thickness is much smaller than the size of the contact region which, in turn, is much smaller than the curvature radii of the contact surfaces, (c) the variation rate of the gap between the cylindrical surfaces in the contact is small, and (d) the contact of actual surfaces can be replaced by a contact of two elastic

half-planes, the fluid rheology is represented as follows (Bair and Winer 1979):

$$\frac{\partial u}{\partial z} = \frac{\tau_L}{\mu} G\left(\frac{\tau}{\tau_L}\right), \quad G(x) = \tanh^{-1} x, \quad (5)$$

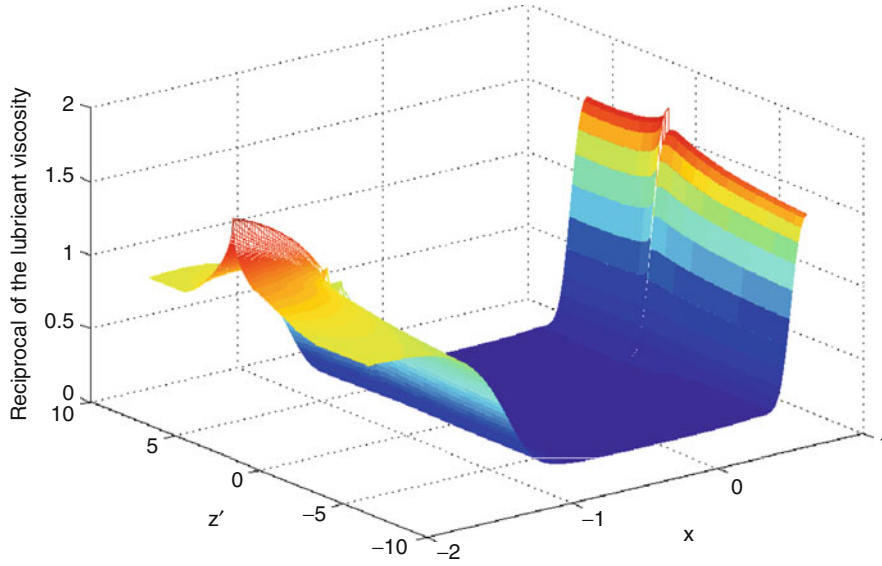
where z is the coordinate of a lubricant particle along the coordinate axis directed across the gap between the cylinders, u is the x -component of the lubricant velocity along the x -axis directed along the direction of the cylinders' motion (and the lubrication layer), τ and τ_L are the local and limiting shear stresses in the lubricant, μ is the lubricant viscosity, $\mu = \mu(x, z)$, G is a given function that determines the rheology of the lubricant fluid, and τ_L is usually a linear function of pressure p , that is,

$$\tau_L = \tau_{L0} + \tau_{L1} p, \quad (6)$$

where parameters τ_{L0} and τ_{L1} are certain functions of the lubricant temperature.

The problem is reduced to a coupled system of the generalized Reynolds equation for non-Newtonian lubricant flow, the equation for the gap between the surfaces of the elastic solids and for the sliding frictional stress in the contact, the equations for the lubricant flow streamlines, the kinetic equation describing the changes in the polymer molecular weight distribution due to degradation, and the equations for the lubricant viscosity (Kudish and Airapetyan 2003, 2004; Kudish and Covitch 2010). The generalized isothermal EHL equations are coupled with the kinetic equation through the lubricant viscosity, which depends not only on lubricant pressure but also on the concentration of polymer molecules and the distribution of their chain lengths. The solution of the problem is obtained using numerical methods similar to the ones described in (Kudish and Airapetyan 2003, 2004; Kudish and Covitch 2010). The kinetic equation is solved along the lubricant flow streamlines. The solution of the kinetic equation predicts the density of the probabilistic distribution of the polymer molecule chain lengths. The shear stress and the changes in the distribution of polymer molecular weight caused by lubricant degradation affect local lubricant properties. The lubricant viscosity experiences reversible and irreversible losses and, in general, is a discontinuous function of spacial variables (see Fig. 4). The changes in the lubricant viscosity alter virtually all parameters of the lubricated contact such as film thickness, friction stresses, pressure, and gap. Several comparisons of lubricants with Newtonian and non-Newtonian rheologies with and without lubricant degradation are considered.

The problem is considered in the following dimensionless variables:



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 4 The reciprocal of the viscosity μ in the lubricants with non-Newtonian and Newtonian rheologies under mixed rolling and sliding conditions and Series I input data ($s_0 = -0.5$). The variable z' is an artificially stretched z -coordinate across the film thickness (namely, $z' = zh(a)/h(x)$) to make the relationship more transparent

$$\begin{aligned}
 (x', a, c) &= \frac{1}{a_H}(x, x_i, x_e), (z', h') = \frac{1}{h_e}(z, h), \\
 p' &= \frac{p}{p_H}, \quad \mu' = \frac{\mu}{\mu_a}, \\
 u' &= \frac{2u}{u_1 + u_2}, \quad w' = \frac{2a_H w}{(u_1 + u_2)h_e}, \\
 (f', \tau'_i) &= \frac{2h_e}{\mu_a(u_1 + u_2)}(f, \tau), \quad \tau'_{L0} = \frac{\tau_{L0}}{p_H}, \\
 (W, W'_a) &= \frac{1}{W_0}(W, W_a),
 \end{aligned} \quad (7)$$

and parameter τ_{L1} and

$$\begin{aligned}
 a &= \frac{x_i}{a_H}, \quad V = \frac{2\pi^2 \mu_a (u_1 + u_2) R' E'}{p^2}, \\
 Q &= \alpha p_H, \quad s_0 = \frac{2(u_2 - u_1)}{u_1 + u_2}, \quad \varepsilon = \left(\frac{a_H}{2R'}\right)^2, \\
 \gamma &= \alpha \frac{U_A}{kT}, \quad \delta = \frac{U_A a_H^2}{Ca_* l_*^2 \mu_a (u_1 + u_2) R'}, \\
 \theta &= k' c_p W_m^\beta, \quad \kappa = \tau_f \frac{u_1 + u_2}{2a_H}.
 \end{aligned} \quad (8)$$

where x_i and x_e are the dimensional inlet and exit coordinates of the contact, $W_a(l)$ is the polymer molecular weight distribution entering the contact, a_H and p_H are the half-width of and the maximum pressure in a dry Hertzian contact, and W_0 is the characteristic value of

the density of molecular weight distribution. For simplicity, in the further discussion primes at the dimensionless variables are omitted.

The solution of the problem is represented by the dimensionless functions: pressure $p(x)$, gap $h(x)$, sliding frictional stress $f(x)$, lubricant viscosity $\mu(x, z)$, and distribution of molecular weight $W(x, z, l)$, and by two dimensionless constants: the exit coordinate c and the exit film thickness:

$$H_0 = \frac{2h_e R'}{a_H^2}. \quad (9)$$

The lubricant flow topology depends on the value of the slide-to-roll ratio s_0 . In case of pure rolling, $s_0 = 0$ and the flow is symmetric about the x -axis while in all other cases it is not. The flow topology determines the way the polymer additive degrades. The detailed analysis of the flow topology is given in (Kudish and Airapetyan 2003, 2004; Kudish and Covitch 2010). The very first step toward the numerical solution of the problem involves choosing an initial approximation. After that the general iterative process is organized as follows. For the known values of pressure $p(x)$, gap $h(x)$, lubricant viscosity $\mu(x, z)$, film thickness H_0 , and exit coordinate c the numerical procedure for solution of the considered problem consists of several steps in the following order: (a) evaluating the sliding frictional stress $f(x)$; (b) determining the

horizontal component of the fluid velocity $u(x, z)$ and fluid flux; (c) calculating the flow streamlines $z(x)$; (d) finding separatrices of the lubricant flow, some of which may be the curves of discontinuity of the lubricant viscosity; (e) evaluating the lubricant viscosity $\mu(x, z)$ at every point of the flow by solving the kinetic equation along the flow streamlines; (f) calculating the new approximation of the sliding frictional stress $f(x)$; and (g) solution of the modified Reynolds and gap equations for $p(x)$, $h(x)$, H_0 , and c . For more information on stable numerical approaches to solution of EHL problems.

Below, two series of results are presented for different combinations of parameters C and α : Series I for $C = 15.5$, $\alpha = 0.008$, and Series II for $C = 0.055$, $\alpha = 0.008$. In Series II simulation of only nominally Newtonian lubricant is considered. In both series of simulations the following values of the dimensional parameters $U = 347$ kJ/mole, $T = 310$ K, $a_* = 0.374$ nm, $l_* = 0.154$ nm, $w_m = 35.1$ g/mole, $a_H = 10^{-3}$ m, $p_H = 1.284$ GPa, and dimensionless parameters $a = -2$, $V = 0.1$, $\varepsilon = 0.01$, $\gamma = 1.0774$, $\kappa = 1$, are used. For Series I simulations, additional values of dimensional parameters (Billmeyer 1966; Crespi et al. 1977) are as follows: $\mu_a = 0.00125$ Pa · s, $c_p = 1.204$ g/dL, and dimensionless parameters $Q = 5$, $\tau_{L0} = 0.002$, $\tau_{L1} = 0.046$, $\delta = 0.671 \cdot 10^8$, $\theta = 4.524 \cdot 10^{-3}$ are used, while for Series II simulations the values of additional dimensional parameters (Billmeyer 1966; Crespi et al. 1977; Kudish and Airapetyan 2003) are $\mu_a = 0.00924$ Pa · s, $c_p = 0.86$ g/dL, and dimensionless parameters $Q = 11$, $\delta = 0.256 \cdot 10^{10}$, $\theta = 3.231 \cdot 10^{-3}$ are used.

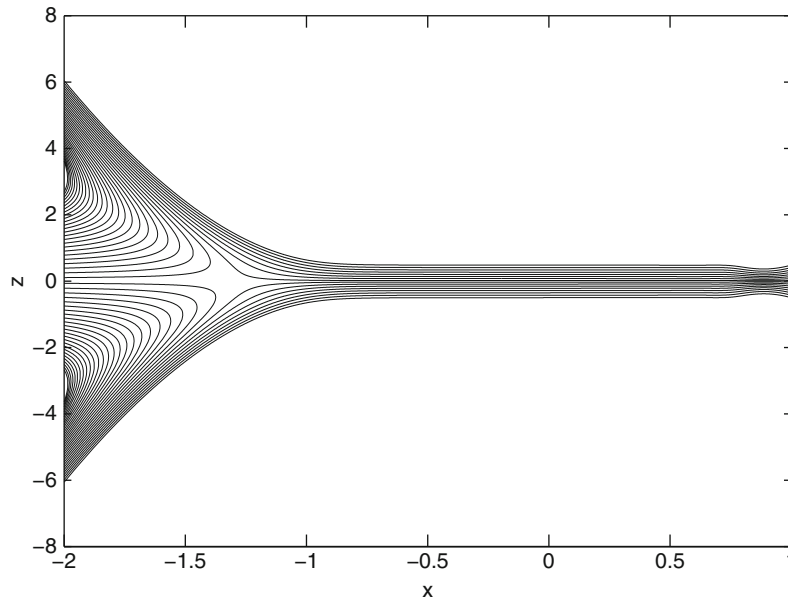
Due to the fact that in the inlet zone pressure p and its derivative dp/dx are relatively small, while gap h is relatively large (and, therefore, the sliding frictional stress f is small), the rheological function G behaves like a linear function (i.e., the lubricant behavior is close to the one of a Newtonian fluid). In heavily loaded EHL contacts film thickness H_0 is primarily determined by the inlet zone (Kudish and Covitch 2010). That explains why in contacts with Newtonian and non-Newtonian lubricants film thicknesses H_0 are close. On the other hand, in the above cases for high slide-to-roll ratios s_0 the sliding frictional stress $f(x)$ for a lubricant with non-Newtonian rheology ($G(x) = \tanh^{-1}x$) is noticeably lower than that for a lubricant with Newtonian rheology ($G(x) = x$) given the same ambient lubricant viscosity.

For a nondegrading lubricant with Newtonian rheology the solution of the isothermal EHL problem is independent of the slide-to-roll ratio s_0 . In particular, for the Newtonian lubricant for the Series I input data $H_0 = 0.1966$ and $c = 1.0513$, while for the Series II input data $H_0 = 0.339$ and $c = 1.052$. For lubricants with

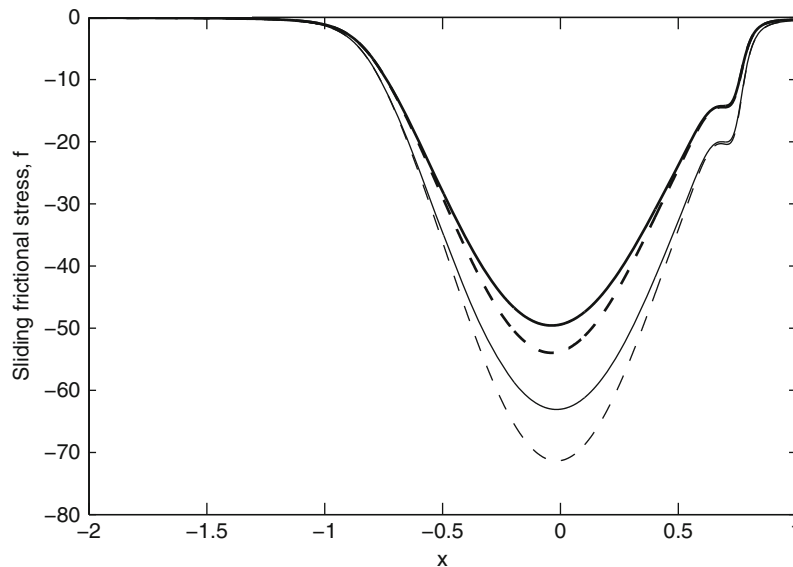
non-Newtonian rheology the solution of the isothermal EHL problem depends on s_0 . For the non-Newtonian lubricant (i.e., for Series I input data), $s_0 = 0$, and τ_{L0} decreasing from 0.01 to 0.002, the film thickness monotonically decreases from $H_0 = 0.1967$ to $H_0 = 0.1963$, respectively, while for $s_0 = -0.5$ the film thickness monotonically decreases from $H_0 = 0.1967$ to $H_0 = 0.1961$, respectively. Therefore, the value of H_0 is almost independent of the slide-to-roll ratio s_0 . The behavior of pressure $p(x)$, gap $h(x)$, film thickness H_0 , coordinate of the exit point c , and the lubricant flow streamlines $z(x)$ in the cases of Newtonian and non-Newtonian lubricants are very similar. For Series I input data and $s_0 = 0$ the graphs of the flow streamlines $z(x)$ for the lubricant with Newtonian rheology are given in Fig. 5. Function $f(x) = 0$ for $s_0 = 0$ while for $s_0 = -0.5$ the graphs of $f(x)$ for the nondegrading Newtonian and non-Newtonian lubricants are presented in Fig. 6. The magnitude of $f(x)$ is greater in the case of the Newtonian lubricant than in the case of the non-Newtonian one. This effect is due to the reversible viscosity loss of the non-Newtonian lubricant in comparison with the Newtonian one. From Fig. 6 it follows that for $s_0 = -0.5$ the reversible lubricant viscosity loss of the non-Newtonian lubricant reaches 12% of its original value. A similar behavior of the surface frictional stresses τ_1 and τ_2 can be seen in Fig. 7, where τ_1 and τ_2 are given for the lubricants with Newtonian and non-Newtonian rheologies under mixed rolling and sliding conditions ($s_0 = -0.5$) as well as the surface shear stress $\tau_1 = -\tau_2$ for the Newtonian lubricant (dashed-dotted line) under pure rolling conditions ($s_0 = 0$). For the case of the pure rolling ($s_0 = 0$) for the non-Newtonian lubricant the surface shear stresses $\tau_1 = -\tau_2$ are very close to the ones for the case of the Newtonian lubricant.

For Series II input data in the case of pure sliding ($s_0 = -2$) the behavior of pressure $p(x)$ and gap $h(x)$ is still very close to the one for the case of pure rolling ($s_0 = 0$). However, the behavior of the streamlines in case of pure sliding ($s_0 = -2$) is different and it is shown in Fig. 8.

Consider some general properties of a solution of the isothermal EHL problem for a degrading lubricant with non-Newtonian rheology. The behavior of pressure $p(x)$ and gap $h(x)$ distributions as well as of the film thickness H_0 and exit coordinate c are practically identical to those in contacts with nondegrading lubricant (Kudish and Covitch 2010). For a lubricant with Newtonian rheology increase in load P leads to increase in the Hertzian pressure p_{Hb} which, in turn, increases Q and causes a relatively slow increase in H_0 and a rapid increase in $f(x)$ and, therefore,



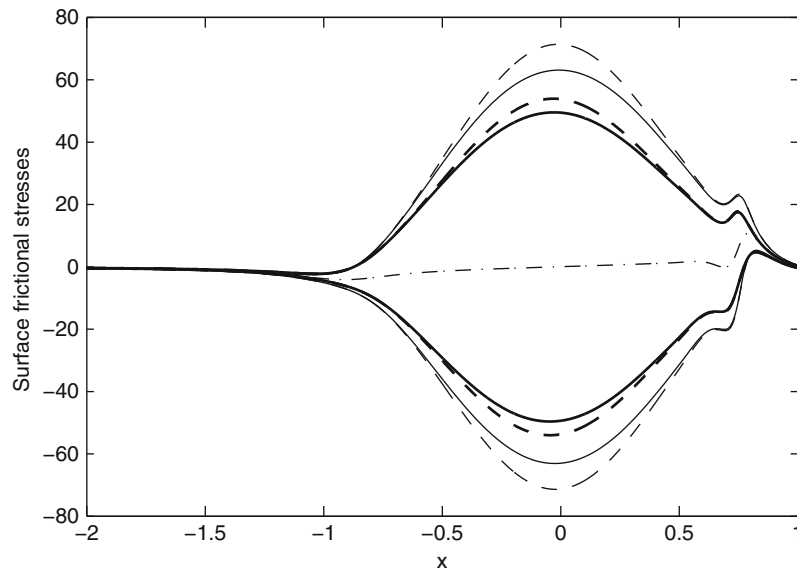
Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 5 Flow streamlines $z(x)$ for nondegrading lubricant with Newtonian rheology and Series I input data, $s_0 = 0$



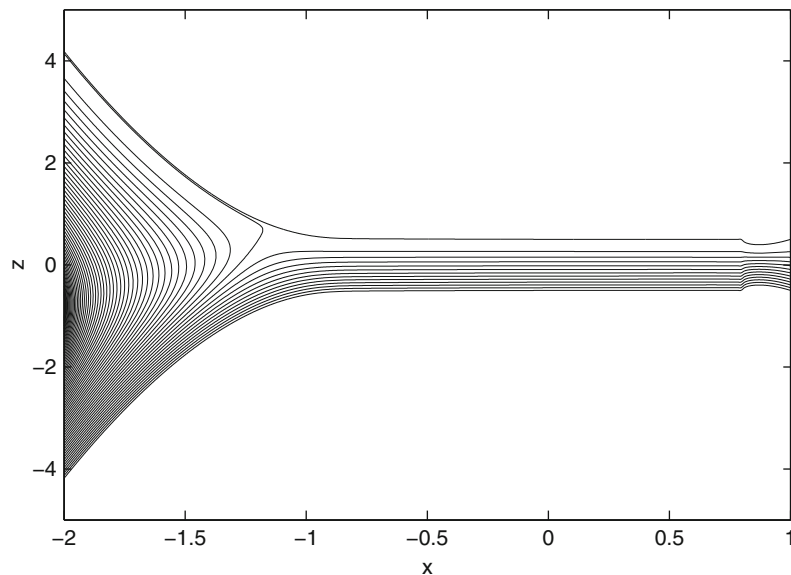
Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 6 Sliding frictional stress $f(x)$ distributions for nondegrading lubricants with Newtonian (*dashed line*) and non-Newtonian (*solid line*) rheologies and for degrading lubricants with Newtonian (*thick dashed line*) and non-Newtonian (*thick solid line*) rheologies for Series I input data, $s_0 = -0.5$

in $\tau(x, z)$. That, in turn, leads to a rapid lubricant degradation. For a lubricant with non-Newtonian rheology, increase in load P also leads to increase in the sliding frictional stress $f(x)$ and, thus, to increase in the shear

stress τ . However, these increases in $f(x)$ and τ are moderated by the lubricant rheology (i.e., these increases are bounded by the limiting stress τ_L). Therefore, for a non-Newtonian lubricant the shear stress τ for high loads P is



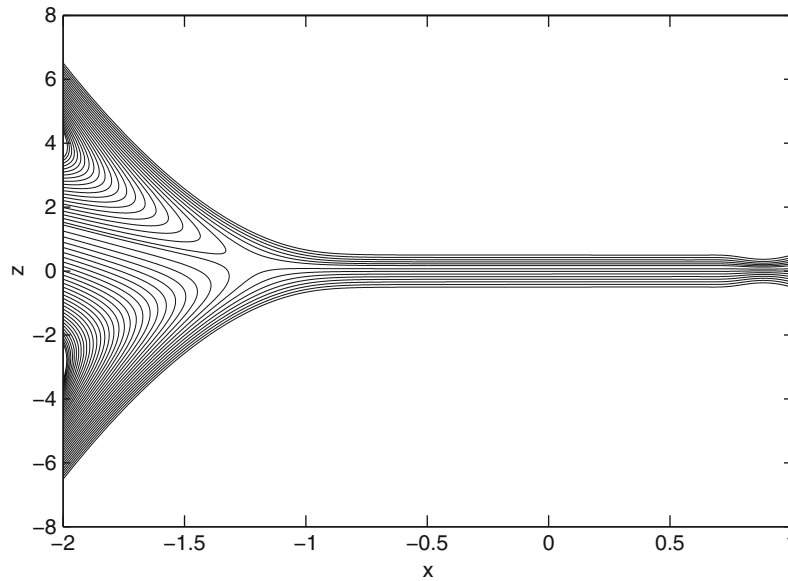
Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 7 Frictional stresses τ_1 and τ_2 applied to the contact surfaces for nondegrading (*dashed lines*) and degrading (*thick dashed lines*) Newtonian lubricants and for nondegrading (*solid lines*) and degrading (*thick solid lines*) non-Newtonian lubricants under mixed rolling and sliding conditions ($s_0 = -0.5$) and surface frictional stresses $\tau_2 = -\tau_1 = (6H_0^2/V)hdp/dx$ for nondegrading Newtonian lubricant (*dash-dotted line*) under pure rolling conditions ($s_0 = 0$) and Series I input data



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 8 Flow streamlines $z(x)$ for nondegrading lubricant with Newtonian rheology and Series II input data, $s_0 = -2$

much lower than for a similar Newtonian lubricant. This means that the degradation process of such a non-Newtonian lubricant runs slower than for a Newtonian counterpart. A usually very small parameter ε almost does

not affect the problem solution. For small τ_{L0} and τ_{L1} the shear stress τ in a non-Newtonian lubricant is small, which, in turn, slows down the process of lubricant degradation (see equation (2) for R and L_*). As the values of



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 9 Flow streamlines for degrading lubricant with non-Newtonian rheology under pure rolling conditions ($s_0 = 0$) and Series I input data

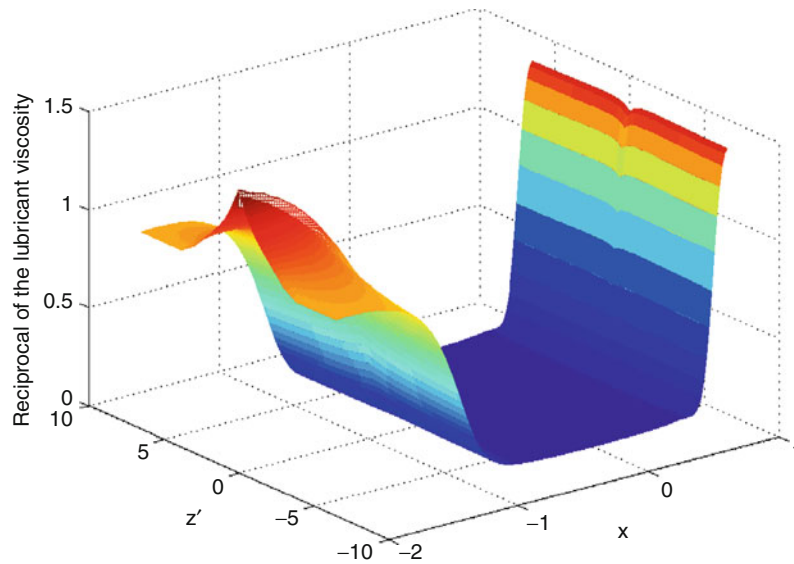
τ_{L0} and τ_{L1} increase the lubricant shear stress τ and the film thickness H_0 increase. As the slide-to-roll ratio s_0 increases, the film thickness H_0 slowly decreases. However, variations in H_0 are insignificant while the increase in τ is directly proportional to the increase in τ_L (i.e., the increase in τ_{L0} and τ_{L1}). The latter causes lubricant degradation to run faster (see equation (2) for R and L_*). Moreover, an increase in the value of parameter δ leads to a corresponding increase in the characteristic chain length L_* that causes lubricant degradation to slow down. For smaller values of the parameter γ the process of lubricant degradation runs slower than for the larger ones. For higher values of θ the lubricant viscosity responds stronger to changes in the molecular weight distribution (i.e., for higher θ the loss of lubricant viscosity is higher). The value of the parameter κ in the left-hand side of the kinetic equation controls the rate of lubricant convection and, therefore, the time a lubricant small volume is present in the contact area. For high values of κ the time of lubricant presence in the contact area is small and the rate of lubricant degradation is low.

Consider the case of pure rolling ($s_0 = 0$) for Series I input parameters. Under these conditions lubricants with Newtonian and non-Newtonian rheologies degrade to a lesser extent than in cases when $s_0 \neq 0$. The solutions for Newtonian and non-Newtonian lubricant rheologies are qualitatively and quantitatively very close to each other

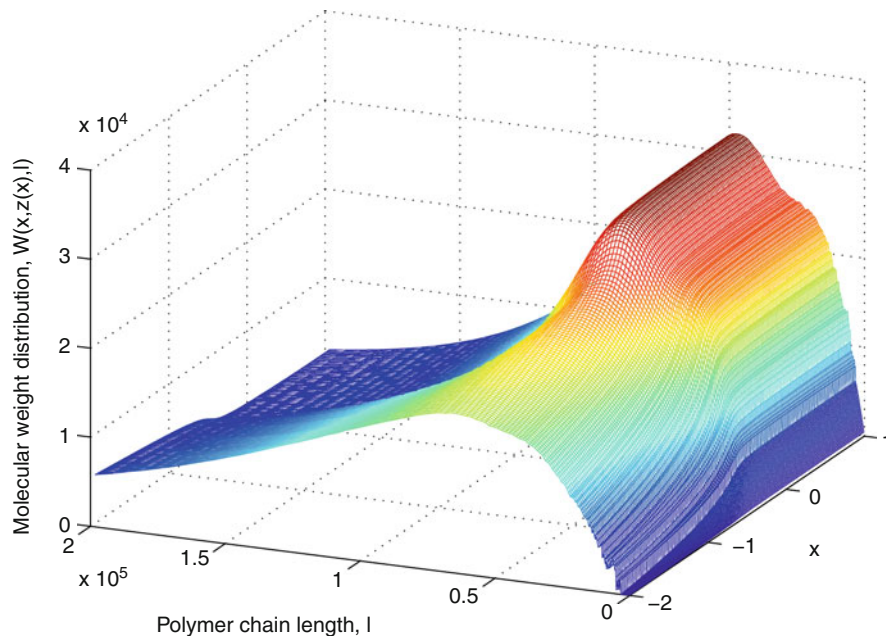
because $|G^{-1}(x)| \leq |x|$ and $G(x) \rightarrow x$ as $x \rightarrow 0$. The map of flow streamlines for the non-Newtonian lubricant is presented in Fig. 9. The pressure and gap distributions are very close to the ones for the nondegrading Newtonian lubricant. The maximum difference between the gap distributions for Newtonian and non-Newtonian fluids reaches 8%.

The maps of the horizontal component u of the lubricant velocity along the flow streamlines are given in (Kudish and Airapetyan 2004; Kudish and Covitch 2010). The exit film thicknesses and exit coordinates for the Newtonian and non-Newtonian degrading lubricants are $H_0 = 0.1829$, $c = 1.0527$ and $H_0 = 0.1834$, and $c = 1.0527$, respectively. The comparison of the data for H_0 obtained for degrading and nondegrading lubricants for pure rolling $s_0 = 0$ shows a relatively small affect of lubricant degradation on the film thickness H_0 (only about 8%).

The distributions of the lubricant viscosity μ and polymer molecular weight W for lubricants with Newtonian and non-Newtonian rheologies are practically identical. The fact that for $s_0 = 0$ lubricants with Newtonian and non-Newtonian rheologies degrade relatively slowly can be also seen from the graphs of the reciprocal of the lubricant viscosity (see Fig. 10) and the distribution of the molecular weight W along the flow streamline $z(x)$ that is closest to the line $z = 0$ and running through the



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 10 The reciprocal of the lubricant viscosity μ in Newtonian and non-Newtonian lubrication film under pure rolling conditions ($s_0 = 0$) and Series I input data. The variable z' is an artificially stretched z -coordinate across the film thickness (namely, $z' = zh(a)/h(x)$) to make the relationship more transparent



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 11 Molecular weight distribution W of degrading lubricants with Newtonian and non-Newtonian rheologies along the flow streamline closest to $z = 0$ and running through the whole contact below $z = 0$ under pure rolling conditions ($s_0 = 0$) and Series I input data

whole contact below $z = 0$ (see Fig. 11). It is clear from the graphs of the reciprocal of the lubricant viscosity (see Fig. 10) that for the Newtonian and non-Newtonian lubricants the maximum irreversible viscosity loss reaches

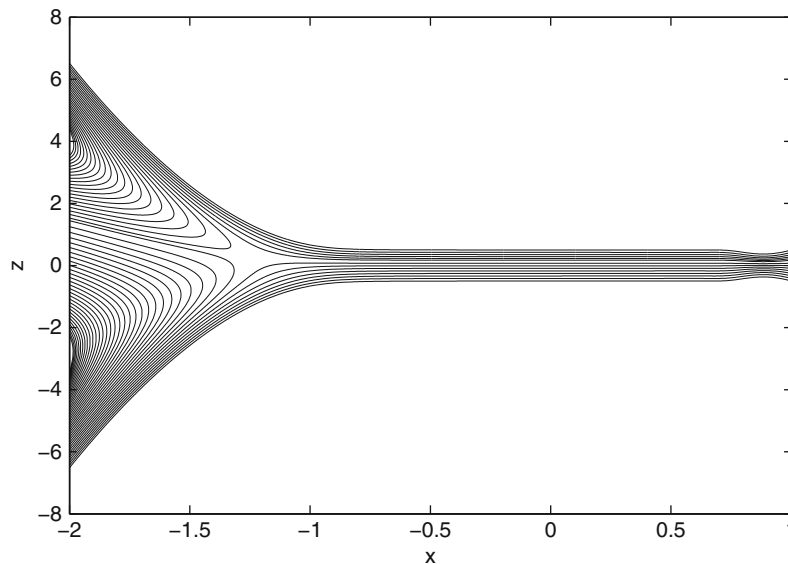
about 31%. Lubricant degradation depends on a number of parameters, among which the shear stress τ plays one of the major roles. For $s_0 = 0$ the sliding frictional stress $f(x) = 0$ and, therefore, the shear stress

$\tau = (12H_0^2/V)zdp/dx$ reaches its extrema in the inlet and exit zones while in the Hertzian region it is relatively small (Kudish and Covitch 2010).

This leads to the conclusion that for pure rolling ($s_0 = 0$) practically all lubricant degradation occurs in the inlet zone, while the lubricant almost does not degrade in the Hertzian region and the exit zone. The latter also can be clearly seen from the distributions of the molecular weight W in Fig. 11.

Consider the case of mixed rolling and sliding with $s_0 = -0.5$ for Series II input data. In this case the solutions of the problem for the lubricants with Newtonian and non-Newtonian rheologies are also very close to each other. The map of the flow streamlines $z(x)$ for the degrading Newtonian and non-Newtonian lubricants is given in Fig. 12. The exit lubrication film thickness H_0 and the exit coordinate c exhibit behavior very similar to the case of pure rolling. For the nondegrading and degrading Newtonian lubricants they are equal to $H_0 = 0.1961$, $c = 1.0513$ and $H_0 = 0.1819$, $c = 1.0531$, respectively. For the nondegrading and degrading non-Newtonian lubricants the film thickness is equal to $H_0 = 0.1961$ and $H_0 = 0.1813$, respectively. In all other respects the distributions of the pressure p and gap h are very similar to the ones for the case of pure rolling. The numerical results show that in the case of mixed rolling and sliding for both lubricants with Newtonian and non-Newtonian rheologies, the sliding frictional

stress $f(x)$ (see Fig. 6) is smaller than in a similar case without degradation but still large enough to cause lubricant degradation to a much greater extent than in the case of pure rolling. In the case of mixed rolling and sliding, the surface frictional stresses τ_1 and τ_2 are much higher (see Fig. 7) than the ones for the case of pure rolling ($\tau_2 = -\tau_1 = (6H_0^2/V)hdp/dx$) and, at the same time, lower than for the case of no degradation. This explains the stronger lubricant degradation in the case of mixed rolling and sliding conditions in comparison with the case of pure rolling conditions (see Figs. 4, 10, 11, and 13). For the degrading lubricants the frictional stresses τ_1 and τ_2 applied to the contact surfaces are on average lower than for the nondegrading lubricant by about 26%. The frictional stress τ in the lubrication film is mostly concentrated near the contact surfaces. Depending on the sign of dp/dx the maximum of the absolute value of the surface frictional stress is reached either on the lower or upper contact surfaces. This leads to faster lubricant degradation near that contact surface and to slower lubricant degradation near the other one. This can be clearly seen from the behavior of the reciprocal of the lubricant viscosity μ in cases of Newtonian and non-Newtonian lubricants (see Fig. 4). In the case of $s_0 = -0.5$ the maximum loss of the lubricant viscosity is approximately 41%. That can be seen from the comparison of graphs of the sliding frictional stress $f(x)$ for the nondegrading and degrading lubricants (see Fig. 6) as well as from the

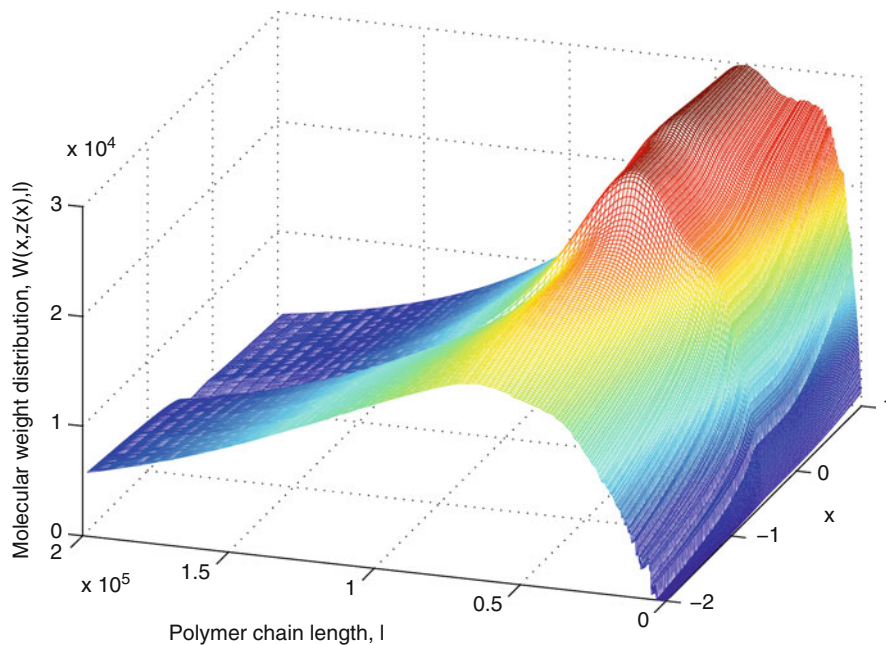


Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 12 Flow streamlines $z(x)$ for degrading lubricants with Newtonian and non-Newtonian rheologies under mixed rolling and sliding conditions ($s_0 = -0.5$) and Series I input data

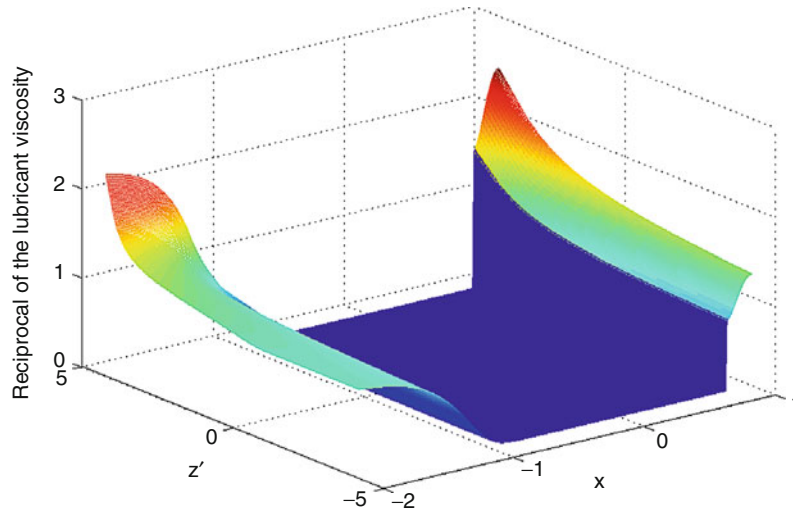
graph of the distribution of the reciprocal of the lubricant viscosity μ (see Fig. 4).

The behavior of the polymer molecule distribution W for the lubricants with Newtonian and non-Newtonian rheologies along one of the flow streamline $z(x)$ running through the whole contact below $z = 0$ and located next to the first turning around streamline is given in Fig. 13. The comparison of the polymer molecular weight W distributions for the cases of pure rolling and mixed rolling and sliding (see Figs. 11 and 13) shows that in the case of pure rolling the lubricant degrades slower than in the case of mixed rolling and sliding. In the latter case rapid polymer scission occurs throughout the entire contact region, while in the former case it is mostly concentrated in the inlet zone of the contact. That can be seen from the values of $W(x, z(x), l)$ for polymer molecules with short chain lengths l at different points along the flow streamlines. Moreover, it follows from Fig. 13 that for Newtonian and non-Newtonian lubricants under mixed rolling and sliding conditions lubricant degradation occurs throughout the contact. For non-Newtonian lubricant at higher values of the slide-to-roll ratio $|s_0|$ the relative impact of the rolling frictional stress $(6H_0^2/V)zdp/dx$ on τ decreases slowly while for the Newtonian lubricant it decreases significantly.

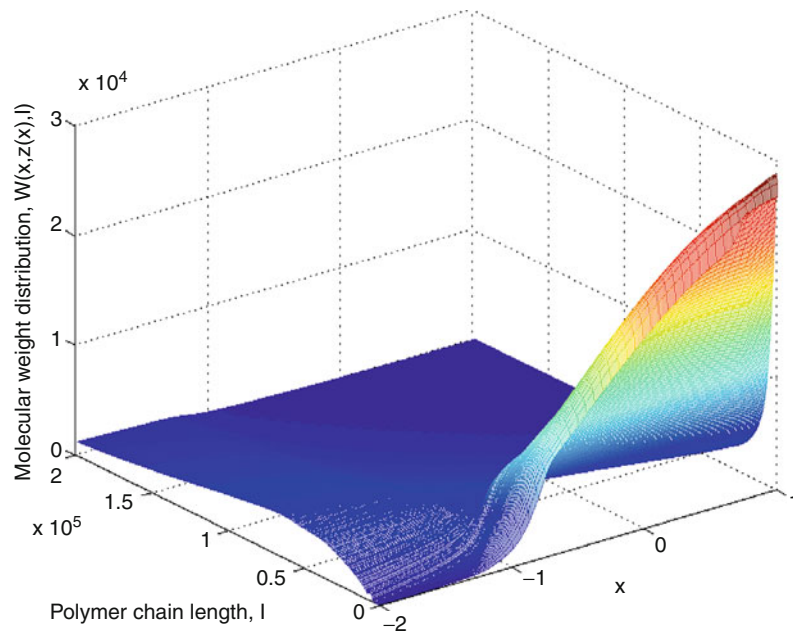
Now, consider the case of Series II input data. Under conditions of pure sliding the behavior of pressure $p(x)$ and gap $h(x)$ in a lubricated contact is very similar to the one described for Series I input data. Even under conditions of pure sliding ($s_0 = -2$) the behavior of pressure $p(x)$ and gap $h(x)$ in a lubricated contact is very similar to the one described for the cases of pure rolling and mixed rolling and sliding. The reduction in the film thickness H_0 due to polymer degradation is slightly higher than for the cases of $s_0 = 0$ and $s_0 = -0.5$, i.e., $H_0 = 0.295$ and $H_0 = 0.339$ with and without lubricant degradation, respectively. For the case of $s_0 = -2$ there is just one set of turning around flow streamlines adjacent to the motionless upper contact surface and one set of flow streamlines running through the whole contact that is adjacent to the moving lower contact surface everywhere in the contact (see Fig. 8). The frictional stresses τ_1 and τ_2 applied to the surfaces are much higher than in the case of pure rolling $s_0 = 0$. The frictional stress τ in the lubrication layer is mostly concentrated near the motionless upper contact surface ($z = h/2$) and it is relatively low near the moving lower contact surface ($z = -h/2$). That leads to significant lubricant degradation near the upper contact surface and to low lubricant degradation near the moving lower contact surface. The numerical results show



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 13 The molecular weight distribution W of degrading lubricants with Newtonian and non-Newtonian rheologies along the flow streamline $z_{31}(x)$ closest to $z = 0$ and running through the whole contact below $z = 0$ under mixed rolling and sliding conditions ($s_0 = -0.5$) and Series I input data



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 14 The reciprocal of the lubricant viscosity μ in a Newtonian lubrication film under pure sliding conditions ($s_0 = -2$) and Series II input data. The variable z' is an artificially stretched z -coordinate across the film thickness (namely, $z' = zh(a)/h(x)$) to make the relationship more transparent



Stress-Induced Lubricant Degradation and Viscosity Loss, Fig. 15 The molecular weight distribution W of the degrading lubricant with Newtonian rheology along the flow streamline $z_{48}(x)$ running through the entire contact closest to the first turning around streamline under pure sliding conditions ($s_0 = -2$) and Series II input data

that, on average, in case of pure sliding a lubricant undergoes degradation to a much greater extent than in the case of pure rolling. This can be clearly seen from the behavior of the lubricant viscosity μ in the lubrication layer (see Fig. 14). In this case the average loss of the

lubricant viscosity reaches about 25%. The local losses of the lubricant viscosity near the moving lower and motionless upper surfaces reach about 10% and 60%, respectively. For the degrading lubricant the frictional stresses applied to the contact surfaces are generally lower than for

Stress-Induced Lubricant Degradation and Viscosity Loss,**Table 1** Fatigue life N versus the survival probability P_{glob} for $s_0 = -0.5$ and $w_m l_{m0} = 1.3196 \cdot 10^5$ ($w_m = 35.1$ g/mol is the monomer weight) as a function of the initial mean chain length of polymer molecules l_m

$P_{glob}(N)$	N	λ	l_m
0.9	$0.149 \cdot 10^{10}$		
0.75	$0.1845 \cdot 10^{10}$	$\lambda = 0.02$	$l_m = l_{m0}$
0.5	$0.2345 \cdot 10^{10}$		
0.9	$0.227 \cdot 10^{10}$		
0.75	$0.2805 \cdot 10^{10}$	$\lambda = 0.01925$	$l_m = 1.25l_{m0}$
0.5	$0.357 \cdot 10^{10}$		
0.9	$0.432 \cdot 10^9$		
0.75	$0.535 \cdot 10^9$	$\lambda = 0.02231$	$l_m = 0.5l_{m0}$
0.5	$0.680 \cdot 10^9$		
0.9	$0.161 \cdot 10^9$		
0.75	$0.200 \cdot 10^9$	$\lambda = 0.02427$	$l_m = 0.25l_{m0}$
0.5	$0.254 \cdot 10^9$		

the nondegrading lubricant by about 20–30%. The general behavior of the polymer molecule distribution $W(x, z(x), l)$ to a certain extent is similar to the one obtained under pure rolling and mixed rolling and sliding conditions and it is given along the flow streamline $z(x)$ running through the entire contact and located next to the first turning around streamline (see Fig. 15). Moreover, in case of pure sliding the lubricant degrades along the flow streamlines throughout the entire contact (see Fig. 15), while in the considered cases of pure rolling and mixed rolling and sliding the degradation occurs mostly in the inlet and the beginning of the Hertzian zones (see Figs. 11 and 13).

It is important to realize that the lubricant viscosity loss caused by degradation is controlled by the contact operating conditions as well as by the lubricant rheology and the nature, concentration, and molecular weight distribution of the polymer additive in the lubricant supplied to the contact. Moreover, significant changes in the lubricant viscosity μ and film thickness H_0 may also lead to noticeable variations in frictional stresses τ_1 and τ_2 applied to the contact surfaces. That affects fatigue life of solids involved in a contact with degrading lubricant (Kudish and Covitch 2010; Kudish 2005) (for more information on contact fatigue modeling see ► [Statistical Fracture Mechanics Approach to Contact Fatigue](#)). An example of such

a dependence of fatigue life N and coefficient of friction λ on the initial mean polymer additive chain length l_m is presented in Table 1.

Therefore, changes in the polymer additive package alone can lead to a 10-fold variation in contact fatigue life.

Key Applications

Modeling of lubricant degradation is important for development of new additives and for predicting useful life of degrading lubricants and their effect on various parameters of lubricated contacts in gears and bearings. The presence of additives in lubricants makes their rheology non-Newtonian. Lubricant degradation is also dependent on polymer additive structure (for example, linear or star polymer structure). Stress-induced lubricant degradation affects lubrication film thickness, frictional stress, wear, and pitting life of gears and bearings. Theoretical prediction of lubricant degradation may allow for better maintenance scheduling of lubricated mechanisms.

Cross-References

- [Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts](#)
- [Numerical Stability and Precision in Elastohydrodynamic Lubrication \(EHL\)](#)
- [Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts](#)
- [Statistical Fracture Mechanics Approach to Contact Fatigue](#)
- [Thermoelastohydrodynamically Lubricated Contacts with Non-Newtonian Lubricants: Asymptotic Approach](#)

References

- S. Bair, W.O. Winer, Shear strength measurements of lubricants at high pressure. *J. Lubr. Tech.* **101**(3), 251–257 (1979)
- F.W. Billmeyer Jr., *Textbook of Polymer Science* (Wiley, New York, 1966)
- M.J. Covitch, How polymer architecture affects permanent viscosity loss of multigrade lubricants. SAE Tech. Paper, No. 982638, 1998
- G. Crespi, A. Valvassori, U. Slisi, Olefin copolymers, in *The Stereo Rubbers*, ed. by W.M. Saltman (Wiley, New York, 1977), pp. 365–431
- H. Eyring, Viscosity, plasticity, and diffusion as examples of absolute reaction rates. *J. Chem. Phys.* **4**(4), 283–291 (1936)
- E. Hoglund, B. Jacobson, Experimental investigations of the shear strength of lubricants subjected to high pressure and temperature. *ASME J. Tribol.* **108**(4), 571–578 (1986)
- I.I. Kudish, Effect of lubricant degradation on contact fatigue. *STLE Tribol. Trans.* **48**(1), 100–107 (2005)

- I.I. Kudish, Modeling of lubricant performance in Kurt Orbahn test for viscosity modifiers based on star polymers. *Int. J. Math. Comp. Model.* **46**(5–6), 632–656 (2007)
- I.I. Kudish, R.G. Airapetyan, Modeling of line contacts with degrading lubricant. *ASME J. Tribol.* **125**(3), 513–522 (2003)
- I.I. Kudish, R.G. Airapetyan, Lubricants with Newtonian and non-Newtonian rheologies and their degradation in line contacts. *ASME J. Tribol.* **126**(1), 112–124 (2004)
- I.I. Kudish, M.J. Covitch, *Modeling and Analytical Methods in Tribology* (Chapman & Hall/CRC, London, 2010)
- I.I. Kudish, R.G. Airapetyan, M.J. Covitch, Modeling of kinetics of strain induced degradation of polymer additives in lubricants. *J. Math. Models Method Appl. Sci.* **12**(6), 835–856 (2002)
- I.I. Kudish, R.G. Airapetyan, M.J. Covitch, Modeling of kinetics of stress induced degradation of polymer additives in lubricants and viscosity loss. *STLE Tribol. Trans.* **46**(1), 1–11 (2003)
- I.I. Kudish, R.G. Airapetyan, G.R. Hayrapetyan, M.J. Covitch, Kinetics approach to modeling of stress induced degradation of lubricants formulated with star polymer additives. *STLE Tribol. Trans.* **48**, 176–189 (2005)
- E.W. Montroll, R. Simha, Theory of depolymerization of long chain molecules. *J. Chem. Phys.* **8**, 721–727 (1940)
- J.A. Odell, A. Keller, Y. Rabin, Flow-induced scission of isolated macromolecules. *J. Chem. Phys.* **88**(6), 4022–4028 (1988) (and references therein)
- O. Saito, On the effect of high energy radiation to polymers I, cross-linking and degradation. *J. Phys. Soc. Jpn.* **13**, 198–206 (1958)
- R.M. Ziff, E.D. McGrady, The kinetics of cluster fragmentation and depolymerization. *J. Phys. A: Math. Gen.* **18**, 3027–3037 (1985)
- R.M. Ziff, E.D. McGrady, Kinetics of polymer degradation. *AchS, Macromol.* **19**, 2513–2519 (1986)

Stress-Induced Polymeric Additive Degradation

► [Stress-Induced Lubricant Degradation and Viscosity Loss](#)

Stress-Life Theories

WEICHENG CUI, FANG WANG
China Ship Scientific Research Center, Wuxi Jiangsu,
People's Republic of China

Synonyms

[S-N curve](#); [S-N diagram](#); [Wöhler curve](#)

Definitions

Material can be induced to fail by many repetitions of stress cycles, and the extreme values of every cycle are

lower than the static strength. The number of stress cycles to failure is called stress life, and those theories used to predict the stress life for a particular structural detail under a specific loading history are called stress-life theories.

To fully describe a stress cycle, two independent quantities among the maximum (σ_{\max}), the minimum (σ_{\min}), the range ($\Delta\sigma = \sigma_{\max} - \sigma_{\min}$), the amplitude ($\sigma_a = \Delta\sigma/2$), the mean value ($\sigma_m = (\sigma_{\max} + \sigma_{\min})/2$), and the ratio ($R = \sigma_{\min}/\sigma_{\max}$) are required from a mathematical point of view. In the earliest series of fatigue tests carried out by Wöhler (e.g., Wöhler 1870), he concluded that the stress amplitudes are decisive for the destruction of the cohesion of the material. The maximum stress is of influence in so far as the higher it is, the lower are the stress amplitudes that lead to failure. Wöhler therefore stated the stress amplitudes to be the most important parameter for fatigue life, but a tensile mean stress also to have a detrimental influence. This conclusion is still used today in many industry design codes, but a study by Vasudevan, Sadananda, and their co-workers (e.g., Vasudevan et al. 2001) indicated that this conclusion is true for most R values, however, for some R values, the maximum stress value could be more important than the stress range. According to their unified approach, the stress life is a function of both stress range and the maximum stress, that is, $N_f = f(\Delta\sigma, \sigma_{\max})$. However, up until now this appeal has not been received adequate attention by industry and more research is needed to demonstrate its necessity.

Scientific Fundamentals

S-N Expressions

The stress-based approach is the earliest and still is the most frequently used approach for fatigue life prediction. Dowling (2007) presented a description on S-N curve. If a test specimen of a material or an engineering component is subjected to a sufficiently severe cyclic stress, a fatigue crack or other damage will develop, leading to complete failure of cyclic stress. If the test is repeated at a higher stress level, the number of cycles to failure will be smaller. The results of such tests from a number of different stress levels may be plotted to obtain a stress versus life curve, also called an S-N curve. In this curve, the fatigue life (number of cycles N) is related to the applied stress range ($\Delta\sigma$ or S) or stress amplitude (σ_a). Wöhler had already carried out experiments to obtain S-N curves in the nineteenth century. A plot of the fatigue life vs. true stress amplitude for a metal gives, in general, a curve of the Basquin form:

$$\sigma_a = \frac{E \cdot \Delta \varepsilon_e}{2} = \sigma'_f \cdot (2N)^b \quad (1)$$

where N is the cycles to failure, $2N$ is load reversals to failure, σ'_f is the fatigue strength coefficient, b is the fatigue strength exponent – the sign of b is negative, $\Delta \varepsilon_e$ is the elastic strain amplitude, and E is the Young's modulus.

For a long time such curves were labeled as Wöhler curves instead of the now more frequently used term S-N curve. In the twentieth century, numerous fatigue tests were carried out to produce large numbers of S-N curves (Schijve 2003). Initially, the fatigue life N was plotted on a logarithmic scale in the horizontal direction, and the stress amplitude on a linear scale in the vertical direction. For low stress amplitudes, the S-N curve exhibited a lower limit, which implies that fatigue failures did not occur after high numbers of load cycles, and the horizontal asymptote of the S-N curve is called the fatigue limit. It is pointed out by Schijve (2003) that cycles with amplitudes larger than the fatigue limit can initiate a fatigue crack, and later cycles with amplitudes below the fatigue limit can propagate the crack and thus become damaging. A simple procedure to account for this phenomenon is to extrapolate the S-N curve below the fatigue limit. A noteworthy proposal was made by Haibach (1970). He started from the idea that the S-N curve is a linear function in a double logarithmic plot, which is known as the Basquin equation, and with the empirical slope factor as the negative inverse slope of the $\log(\sigma_a)/\log N$ plot. Haibach proposed to extend the S-N curve with a second linear part, as shown in Fig. 1, with a slope depending on k of the first part (slope factor $k' = 2k - 1$). Another suggestion to extend the S-N curve was made for welded joints by Mott (1958), as also indicated in Fig. 1.

In a component or structure, there are two types of stress concentration. One is due to the structural geometry

change or discontinuity and the other is due to welding. Depending on how to account for the stress concentration effect, stress-based approaches can be further divided into nominal stress approach, hot-spot stress approach, and notch stress approach.

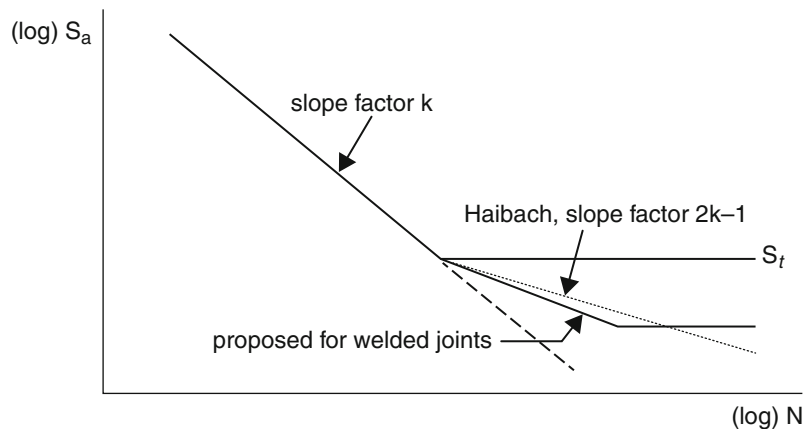
Traditional fatigue tests indicated a limit at about $N = 10^7$ cycles. If the applied stress range is lower than the limit, there will be no fatigue failure. However, recent gigacycle fatigue tests (e.g., Wang et al. 1999) showed that the fatigue limit does not appear until 10^9 cycles in the S-N curve or S-N curve tends to come down again in the long life region of $N > 10^7$ cycles. This poses questions on the existence of infinite fatigue life. Thus, whether a fatigue limit exists requires further study.

If a fatigue limit is assumed to exist, then the relation expressed by (1) is only valid for the middle part. The stress amplitude is larger than the limit but the maximum stress should not exceed the ultimate tensile strength. A new function for the description of fatigue curves for both low and high fatigue regions, that is, for the whole cycle region from tensile strength to fatigue limit, has been proposed by Kohout and Vechet (1999). The function takes the following form:

$$\sigma(N) = \sigma_\infty \left[\frac{N + 10^7 \alpha \beta}{N + 10^7 \alpha} \right]^{-b} \quad (2)$$

$$\alpha = \frac{\sigma_c^{-1/b} - \sigma_\infty^{-1/b}}{\sigma_u^{-1/b} - \sigma_c^{-1/b}} \quad \beta = \frac{\sigma_u^{-1/b}}{\sigma_\infty^{-1/b}}$$

where σ_u is the tensile strength and σ_∞ is the fatigue limit; both can be measured accurately. σ_c is the fatigue strength at 10^7 cycles and $-b$ is the slope in the middle cycle region. σ_c and b can be determined by the least squares method.



Stress-Life Theories, Fig. 1 Representation of S-N curve extrapolated below the fatigue limit (Schijve 2003)

Four parameters have to be determined for a complete S-N curve according to (2). In some practical situations, not as much information may be available. Wang et al. (1999) proposed a simple and practical prediction method to estimate both the S-N curve and the crack growth rate curve using only the tensile strength.

Generally, the basic fatigue behavior of materials and structures is determined by adopting constant amplitude loading resulting in various characteristic quantities and relationships such as S-N curves, cyclic stress-strain curves, fatigue crack growth rates, threshold values, and so on. While constant amplitude loading is fully described by quite a few parameters (maximum/minimum or range/mean and number of load cycles), most structures experience in-service loading environments with variable amplitudes and mean loads. It is well established that fatigue response may be very sensitive to the specifics of the loading encountered.

Therefore, tests with realistic load sequences are often required in order to assess any susceptibility to that kind of phenomenon and to demonstrate the in-service integrity for given materials and structures. For this purpose, standardized load-time histories (SLH) have been studied for 30 years since it has been recognized that the use of SLH provides a series of advantages – both for studies of a more generic nature and practical applications.

Mean Stress Effect

Mechanical components undergoing in-service fatigue loadings exhibit a fatigue strength that generally depends on the mean values of the stress components. A tensile mean normal stress has a detrimental effect on fatigue strength, whereas, in general, a compressive mean normal stress has a beneficial effect on it. Conversely, a non-zero mean shear stress does not affect fatigue strength as long as the applied maximum shear stress is lower than the material shear yield stress (Susmel et al. 2005). The S-N curves are quite different for various mean stresses.

When a component is in fatigue limit conditions, the most employed methods correlate the amplitude and the mean value of the applied cyclic stress by using either the yield or the ultimate stress and the fully reversed plain fatigue limit. A number of empirical relationships capable of accounting for the mean stress effect on the material fatigue strength under uniaxial fatigue loading exist. Susmele et al. (2005) presented a brief summary on the empirical relationships. Some of those expressions could be summarized by Marin's relationship (1956),

$$\left(\frac{\sigma_a}{\sigma_{ar}}\right)^n + \left(f \frac{\sigma_m}{\sigma_u}\right)^m = 1 \quad (3)$$

where σ_a is the nominal stress amplitudes, σ_{ar} is the fully reversed stress amplitude, σ_m is the nominal mean stress, σ_u is the ultimate tensile strength, and f , m , and n are constants to be defined. In particular, the expressions include those proposed by Gerber, Dietman, Goodman, Soderberg, and the so-called “elliptical relationship.”

1. Soderberg's relationship when $n = 1$, $m = 1$, $f = \sigma_u / \sigma_Y$

$$\frac{\sigma_a}{\sigma_{ar}} + \frac{\sigma_m}{\sigma_Y} = 1 \quad (4)$$

where σ_Y is the yield strength of the material.

2. Goodman's relationship when $n = 1$, $m = 1$, $f = 1$

$$\frac{\sigma_a}{\sigma_{ar}} + \frac{\sigma_m}{\sigma_u} = 1 \quad (5)$$

3. Gerber's parabola when $n = 2$, $m = 2$, $f = 1$

$$\frac{\sigma_a}{\sigma_{ar}} + \left(\frac{\sigma_m}{\sigma_u}\right)^2 = 1 \quad (6)$$

4. Dietman's parabola when $n = 2$, $m = 1$, $f = 1$

$$\left(\frac{\sigma_a}{\sigma_{ar}}\right)^2 + \frac{\sigma_m}{\sigma_u} = 1 \quad (7)$$

5. The so-called “elliptical relationship” when $n = 2$, $m = 2$, $f = 1$

$$\left(\frac{\sigma_a}{\sigma_{ar}}\right)^2 + \left(\frac{\sigma_m}{\sigma_u}\right)^2 = 1 \quad (8)$$

Improved agreement for ductile material is often possible by replacing σ_u in (5) with either: (a) the corrected true fracture strength $\tilde{\sigma}_{fB}$ or (b) the constant σ'_f from the unnotched axial S-N curve for $\sigma_m = 0$ (Dowling 2007). The corresponding equations are

$$\frac{\sigma_a}{\sigma_{ar}} + \frac{\sigma_m}{\tilde{\sigma}_{fB}} = 1, \frac{\sigma_a}{\sigma_{ar}} + \frac{\sigma_m}{\sigma'_f} = 1 \quad (9)$$

Such modification of Goodman's relationship was proposed by Morrow. The two values of $\tilde{\sigma}_{fB}$ and σ'_f are somewhat higher than σ_u for ductile metals.

Another expression that is frequently employed is the Smith, Watson, and Topper (SWT) equation, which has the advantage of not relying on any material constant:

$$\sigma_{ar} = \sqrt{\sigma_{\max} \sigma_a} \quad (\sigma_{\max} > 0) \quad \text{or} \quad \sigma_{ar} = \sigma_{\max} \sqrt{\frac{1-R}{2}} \quad (\sigma_{\max} > 0) \quad (10)$$

where $\sigma_{\max} = \sigma_m + \sigma_a$.

An additional one is Walker's expression, which employs a materials constant γ to the relationship,

$$\begin{aligned}\sigma_{ar} &= \sigma_{\max}^{1-\gamma} \sigma_a^{1-\gamma} (\sigma_{\max} > 0) \quad \text{or} \\ \sigma_{ar} &= \sigma_{\max} \left(\frac{1-R}{2} \right)^\gamma (\sigma_{\max} > 0)\end{aligned}\quad (11)$$

γ is obtained by fitting of data for more than one mean stress or R value.

Based on the above expressions representing the amplitude-mean behavior, the life estimates and S-N curves with mean stress effect included can be achieved. σ_{ar} can be regarded as an equivalent completely reversed stress amplitude, hence, the life estimate for the condition with σ_m and σ_a in combination can be obtained by substituting σ_{ar} into an S-N curve for zero mean stress.

Multiaxial Stress Effect

In engineering components, cyclic loadings that cause complex states of stress are common. And different sources may differ in phase or frequency or both (Dowling 2007).

Generally, there are two approaches to consider the effect of multiaxial stress. One is the effective stress

approach; another is the critical plane approach. The first approach considers the simple situation where all cyclic loads are completely reversed and have the same frequency, and further they are either in-phase or 180° out of phase with one another. The amplitudes of the principal stresses σ_{1a} , σ_{2a} , and σ_{3a} can then be employed to compute the effective stress amplitude $\bar{\sigma}_a$ using the following relationship:

$$\bar{\sigma}_a = \frac{1}{\sqrt{2}} \sqrt{(\sigma_{1a} - \sigma_{2a})^2 + (\sigma_{2a} - \sigma_{3a})^2 + (\sigma_{1a} - \sigma_{3a})^2} \quad (12)$$

In applying (12), amplitudes considered to be in-phase are positive, and those 180° out-of-phase are negative.

Then the life can be estimated by using the effective stress amplitude $\bar{\sigma}_a$ into an S-N curve for completely reversed uniaxial stress.

If the mean stress should be considered, then values of effective stress amplitude and mean stress can be calculated from:

$$\bar{\sigma}_a = \frac{1}{\sqrt{2}} \sqrt{(\sigma_{xa} - \sigma_{ya})^2 + (\sigma_{ya} - \sigma_{za})^2 + (\sigma_{xa} - \sigma_{za})^2 + 6(\tau_{xya}^2 + \tau_{yza}^2 + \tau_{zxa}^2)} \quad (13)$$

$$\bar{\sigma}_m = \sigma_{xm} + \sigma_{ym} + \sigma_{zm} \quad (14)$$

Substituting the above two equations to the methods introduced in (4)–(11), the equivalent completely reversed uniaxial stress with mean stress effect can be obtained. Then the life can be estimated by using this equivalent completely reversed uniaxial stress.

For situations where the principal axes rotate during cyclic loading, or cyclic loads occur at more than one frequency, and if there is a difference in phase between them, the second approach, namely the critical plane approach, should be used. Some of the most successful criteria are based on the critical plane approach. In general, these methods are based on the combined use of the shear stress amplitude acting on the plane experiencing the maximum shear stress amplitude (critical plane) and the maximum stress normal to this plane. One of the most important peculiarities of the critical plane approach is that the combined use of the shear stress amplitude, τ_a , and the maximum normal stress, $\sigma_{n,\max}$, relative to the

critical plane permits the influence of both out-of-phase angles and mean stresses to be taken into account easily (Susmel et al. 2005).

Under uniaxial fatigue loadings, it is easy to demonstrate that the stress components relative to the critical plane are

$$\tau_a = \frac{\sigma_{x,\max}}{4} (1-R) = \frac{\sigma_{x,a}}{2} \quad (15)$$

$$\sigma_{n,\max} = \frac{\sigma_{x,\max}}{2} = \frac{\sigma_{x,m} + \sigma_{x,a}}{2} \quad (16)$$

where R is the nominal load ratio.

By combining (15)–(16) with the equations introduced in the part of “mean stress effect,” the idea for uniaxial fatigue can be extended to more complex multiaxial fatigue problem. For example, by substituting (15)–(16) into (3), the Marin’s general relationship can be rewritten as

$$\left(\frac{2\tau_a}{\sigma_{ar}}\right)^n + \left(2f\frac{\sigma_{n,\max} - \tau_a}{\sigma_u}\right)^m = 1 \tag{17}$$

Simplifications for S-N Curve Parameters

When the fatigue properties are evaluated, the S-N curve parameters should be available (e.g., Roessle and Fatemi 2000; Li et al. 2009). Without doubt, the best way to know these parameters is to test. However, fatigue testing is a time-consuming process. Hence, in the absence of experimentally determined values, it is highly desirable to estimate the S-N curve from a knowledge of the more readily available material monotonic properties obtained in simple tensile tests. If reliable corrections with reasonable accuracy can be established, durability performance predictions and optimization analysis can be performed with a substantial reduction of time and cost associated with material fatigue testing (Li et al. 2009).

1. Estimation of fatigue limit σ_{-1}

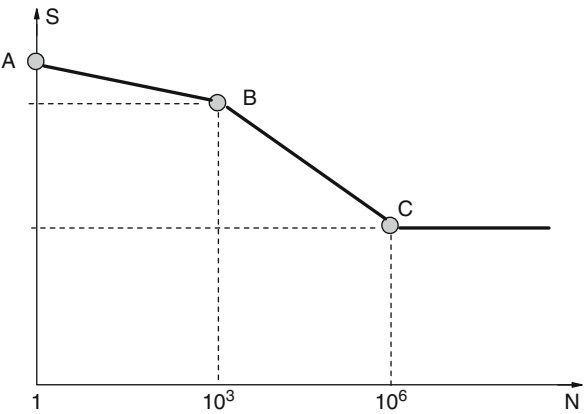
A commonly used approximation of fatigue limit, σ_{∞} , from ultimate tensile strength, σ_u , for low and medium strength carbon and alloy steels was represented by Grover et al. in 1954. Bathias and Bailon (1981) proposed another approximation often used for steels from the ultimate tensile strength and area reduction. Li (2009) further checked the accuracy of the two methods and found that the test data scatter at a quite large range and the predictions based on the former two methods have poor agreement with the test data and are non-conservative for most of the alloy steels. Actually, most traditional fatigue limit data are obtained from rotating bending fatigue tests, whereas the values are from axial fatigue tests (Roessle and Fatemi 2000). In axial fatigue tests, lower strengths are often obtained due to reasons such as higher probability of defects associated with larger stressed volume and bending resulting from specimen misalignment (Roessle and Fatemi 2000). Hence, they developed two new fatigue limit estimation equations for the conditions where the cyclic fatigue limit is known or ultimate tensile strength and area reduction are known, respectively. Error analysis proves the accuracy of the new equations. The methods and their corresponding equations are listed in Table 1.

Combinations of effects such as the type of loading, sample size, and surface finish are common in engineering situations. What is usually done is to multiply reduction factors for the various effects to obtain an adjusted lower fatigue limit. Details can be referred to the book by Dowling (2007).

2. Estimation of S-N curve

Stress-Life Theories, Table 1 Methods to estimate the fatigue limit

Name	Equation	Error
Grover et al. (1954)	$\sigma_{\infty} = 0.5 \cdot \sigma_u (\sigma_u < 1,800 \text{ MPa})$	37.0%
Bathias and Bailon (1981)	$\sigma_{\infty} = 0.39 \cdot \sigma_u + 100\Phi$ Φ : area reduction	24.4%
Li et al. (2009)	$\sigma_{\infty} = 1.13 \cdot \sigma_Y^{0.9}$ σ_Y' : cyclic yield strength	8.1%
Li et al. (2009)	$\sigma_{\infty} = 0.46 \cdot \left(\frac{(1+\Phi)\sigma_u}{(-\ln(1-\Phi))^{0.16}}\right)^{0.9}$	10.0%



Stress-Life Theories, Fig. 2 Estimating completely reversed S-N curve according to procedures

Estimates of fatigue limits can be used as part of a procedure for estimating entire S-N curves. The widely used methods are established corresponding to completely reversed loading, that is, to zero mean stress. The S-N curve is established by three points illustrated in Fig. 2. The positions of the critical three points depend on load types, size (stress gradient), surface finish, and so on. Additional information can be consulted in Dowling (2007) and its reference books.

Key Applications

The stress-based approach is widely applied in fatigue life assessment of engineering structures. In some cases, S-N curves may be available from testing of members very similar to the actual component of interest. Examples might be welded joints, built-up riveted beams, or vehicle axles, of a specific design and material (Dowling 2007). Generally, the basic fatigue behavior of materials and

structures is determined by adopting constant amplitude loading resulting in various characteristic quantities and relationships such as S-N curves, cyclic stress-strain curves, fatigue crack growth rates, threshold values, and so on. While constant amplitude loading is fully described by quite a few parameters (maximum/minimum or range/mean and number of load cycles), most structures experience in-service loading environments with variable amplitudes and mean loads, and fatigue response may be very sensitive to the specifics of the loading encountered.

The applications of stress-based approach have been developed well during its application in real structural details such as non-continuous welded joints. The S-N diagrams for non-continuous welded joints must be adjusted to take into account the effects of differences in geometry between the test specimen underlying the S-N curve and the hull structural detail it is applied to, considering the inelastic response of material at the anticipated crack origin, multiple modes of loading, and their statistical correlation (Petinoy and Thayamballi 1998). At the same time, finite element (FE) analysis is being used by designers for fatigue assessment of structures. The proper link between calculated hot spot stress and fatigue capacity can be established. The fatigue capacity may then be expressed as a hot spot stress S-N curve. The hot spot-stress-based design S-N curve was widely used in practices such as fatigue analysis of welded ship structures. For example, static loads on ship structures induced by cargo loading cause relatively higher stress histories at welded joints compared with cyclic loads induced by waves. Due to these static loads, the initial tensile residual stresses at welded joints are shaken down to a great extent by the elastoplastic deformation behavior of the material. The redistribution of initial welding residual stresses by the preload can be investigated by FE analysis combined with tests to obtain empirical formula of S-N curves, taking into account the effect of the arbitrary preload and mean stress associated with static loads based on the hot-spot stress range (Kang and Kim 2003).

Although stress-based approaches are widely used in industry due to their simplicity, the limitations mentioned in the entry on ► [damage accumulation](#) exist in all these approaches and a way forward is to use fatigue crack propagation theory.

Cross-References

- [Damage Accumulation](#)
- [Fatigue](#)
- [Fatigue Limit](#)

References

- C. Bathias, J.P. Bailon, *La Fatigue des Matériaux et des Structures* (Les Presses D'Université De Montreal, Montreal, 1981) (in French)
- N.E. Dowling, *Mechanical Behavior of Materials-Engineering Methods for Deformation, Fracture, and Fatigue*, 3rd edn. (Pearson Prentice Hall, Upper Saddle River, 2007)
- H.J. Grover, A. Gordon, L.R. Jackson, *Fatigue of Metals and Structures* (US Government Printing Office, Washington, DC, 1954)
- E. Haibach, *Modified Linear Damage Accumulation Hypothesis Accounting for A Decreasing Fatigue Strength During Increasing Fatigue Damage*, Report TM Nr. 50 (Laboratorium für Betriebsfestigkeit, LBF, Darmstadt, 1970) (in German)
- S.W. Kang, W.S. Kim, A proposed S-N Curve for welded ship structures. *Weld J.* **82**(7), 161–169 (2003)
- J. Kohout, S. Vechet, New functions for description of fatigue curves and their advantages, in *Fatigue'99*, (Higher Education Press, Beijing, 1999), pp. 783–788
- J. Li, Q. Sun, Z.P. Zhang, C.W. Li, Y.J. Qiao, Theoretical estimation to the cyclic yield strength and fatigue limit for alloy steels. *Mech. Res. Commun.* **36**(3), 316–321 (2009)
- J. Marin, Interpretation of fatigue strength for combined stresses, in *Proceedings of International Conference Fatigue Metals*, London, 1956, pp. 184–192
- N.F. Mott, A theory of the origin of fatigue cracks. *Acta Metall.* **6**, 195–197 (1958)
- S.V. Petinoy, A.K. Thayamballi, The application of S-N curves considering mismatch of stress concentration between test specimen and structure. *J. Ship Res.* **42**(1), 68–78 (1998)
- M.L. Roessle, A. Fatemi, Strain-controlled fatigue properties of steels and some simple approximations. *Int. J. Fatigue* **22**, 495–511 (2000)
- J. Schijve, Fatigue of structures and materials in the 20th century and the state of the art. *Int. J. Fatigue* **25**, 679–702 (2003)
- L. Susmel, R. Tovo, P. Lazzarin, The mean stress effect on the high-cycle fatigue strength from a multiaxial fatigue point of view. *Int. J. Fatigue* **27**, 928–943 (2005)
- A.K. Vasudevan, K. Sadananda, G. Glinka, Critical parameters for fatigue damage. *Int. J. Fatigue* **23**, S39–S53 (2001)
- Q.Y. Wang, Z.D. Sun, C. Bathias, J.Y. Berard, S. Rathery, Fatigue crack initiation and growth behavior of a thin steel sheet at ultrasonic frequency, in *Fatigue'99*, (Higher Education Press, Beijing, 1999), pp. 169–174
- A. Wöhler, *Über die Festigkeits-Versuche mit Eisen und Stahl, Auf Anordnung des Ministers für Handel, Gewerbe u. öffentl. Arbeiten, Grafen Itzenplitz, angestellt* (Ernst und Korn, Berlin, 1870)

Stribeck Curves

YANSONG WANG¹, Q. JANE WANG²

¹Baker Hughes Incorporated, Houston, TX, USA

²Department of Mechanical Engineering and Center for Surface Engineering and Tribology, Northwestern University, Evanston, IL, USA

Synonyms

[Friction in different lubrication regimes](#); [Lubrication](#); [Variation of friction with operation parameters](#)

Definition

The Stribeck curve is an overall view of friction variation in the entire range of lubrication, including the hydrodynamic, mixed, and boundary lubrication regimes.

Scientific Fundamentals

Machine elements may experience a wide range of lubrication regimes, including full-film, mixed, and boundary lubrication, depending on their operating conditions and lubricant properties, where friction at the interface of components varies. The Stribeck curve is an overall view of friction variation in the entire range of lubrication. In 1902, Professor Richard Stribeck (1902) confirmed the existence of a minimum friction through his extensive journal bearing friction experiments. In 1914, Ludwig Gumbel summarized the Stribeck results in a single curve by means of dimensionless parameters. In the same year, Mayo Hersey (1914) showed that friction due to viscous shear was a unique function of the product of viscosity (η) by rotational speed (N) divided by the average load (P), which is called the Hersey number, $\eta N/P$. The friction coefficient plotted as a function of the Hersey number is now commonly known as the Stribeck curve or Lambda curve. Figure 1 shows a schematic Stribeck curve for a journal bearing system, where regimes of lubrication, the full film, mixed, and boundary film lubrications, are also illustrated.

Friction in the boundary lubrication regime is mainly that due to rubbing of surfaces with some boundary layers, that in the full-film lubrication regime is mainly viscous dissipation, while that in the mixed lubrication regime can be the combination of the two. Friction due to viscous dissipation should be determined from the viscous shear

of the hydrodynamic actions. When surface roughness is considered, Patir and Cheng's average shear stress expression can be used (Patir and Cheng, 1978). The mean viscous shear can be expressed as

$$\bar{\tau} = -\frac{\eta U}{h}(\phi_f \pm \phi_{fs}) \pm \phi_{fp} \quad (1)$$

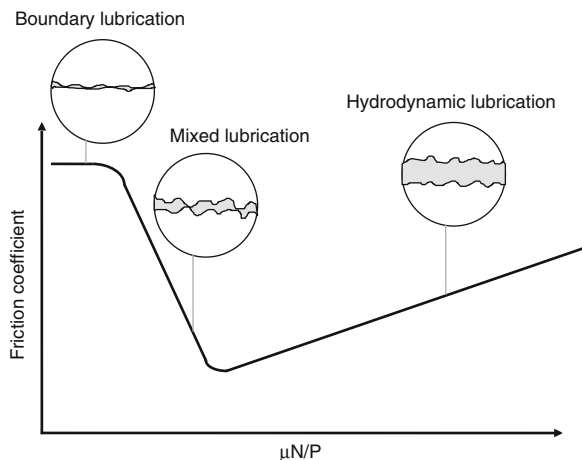
Here, the plus and minus signs refer to the surface at the upper or lower boundaries of the tribological interface. The correction factor for roughness ϕ_f and empirical shear stress factors ϕ_{fs} and ϕ_{fp} are expressed as follows.

$$\begin{cases} \phi_f = h \int_{-h+\delta/100}^{\infty} \frac{g(\delta)d\delta}{h+\delta} \\ \phi_{fp} = 1 - De^{-sH} \\ \phi_{fs} = AH^{\alpha_1} e^{-\alpha_2 H + \alpha_3 H^2} & 0.5 < H \leq 7 \\ \phi_{fs} = 0 & H > 7 \end{cases} \quad (2)$$

where $g(\delta)$ is the density frequency of the combined roughness, $H = h/R_q$ is the non-dimensional film thickness, and D and s are listed in Table 1, while coefficients A , α_1 , α_2 , α_3 are listed in Table 2 as functions of γ , the asperity aspect ratio, or the Peklenik Number (Peklenik 1967).

The friction in the mixed lubrication regime is the summation of the friction due to viscous shear and that due to asperity contact and sliding specified in the boundary lubrication. The total shear stress on a nominal differential area, dA , can be expressed as follows with the boundary friction coefficient f_o related to the asperity contact portion of the area, dA_o , and fluid shear portion of the area, dA_f .

$$\tau_{total} = \frac{\tau dA_f + f_o P_c dA_o}{dA_f + dA_o} \quad (3)$$



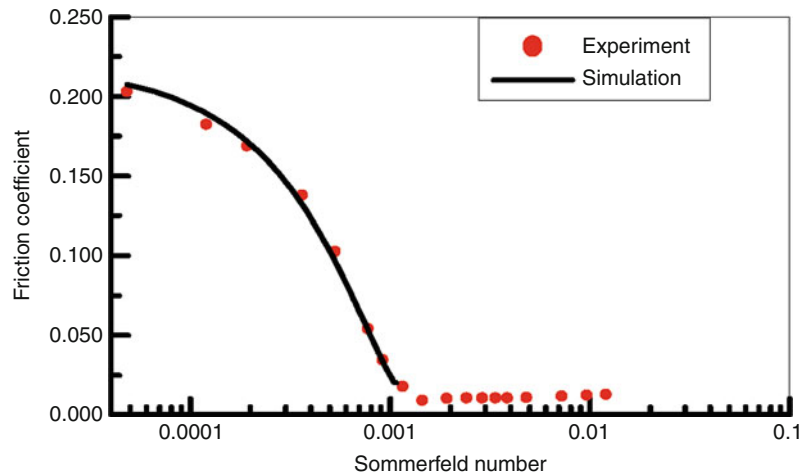
Stribeck Curves, Fig. 1 Schematic of Stribeck curve

Stribeck Curves, Table 1 Coefficients for ϕ_{fp}

γ	D	s	Range
1/9	1.51	0.52	$H > 1$
1	1.40	0.66	$H > 0.75$
9	0.73	0.91	$H > 0.5$

Stribeck Curves, Table 2 Coefficients for ϕ_{fs}

γ	A	α_1	α_2	α_3
1/9	14.1	2.45	2.30	0.10
1	11.1	2.31	2.38	0.11
9	8.7	2.15	2.97	0.18



Stribeck Curves, Fig. 2 Stribeck Curve obtained from an experiment measurement, compared with simulation data (Lu et al. 2006)

In the asperity contact and rubbing area, $dA_f = 0$, while in full-film lubrication regime, $dA_c = 0$. The total friction is integration of (3) over the entire nominal area and the overall friction coefficient is the total friction over this area of interaction divided by the normal load.

Many have conducted simulations and experiments to obtain Stribeck curves; the one by Lu and Khonsari (2006) is an example where the experimental results well agree with theoretical predictions, as shown in Fig. 2, for the coefficient of friction plotted versus a logarithm abscissa of the Sommerfeld number.

The non-dimensional Hersey number mainly concerns operating parameters. It would be better if material properties, surface roughness, and bearing structure, not just fluid viscosity, can also be considered in the evaluation of friction variation, with which the Stribeck concept can be extended to the Stribeck surface. Research on the frictional behavior of a journal bearing conformal contact system through numerical simulations (Wang et al. 2006, Akbarzadeh and Khonsari 2010) explored the possibility of three-dimensional Stribeck surfaces that take into account the effects of bearing structure, heat-transfer conditions, and surface roughness.

The Hersey number is the product of viscosity and rotational speed over the applied load, therefore, the Hersey number can be composed by speed change only without invoking the effect of load. However, load directly influences deformation, and the latter contributes to film thickness. The same Hersey number can be constructed with different combinations of speed and load, and

different deformations, or different film thickness, may occur under different loads, resulting in different friction conditions. There are other parameters affecting the film thickness. This calls for adding a third dimension to the Stribeck curve to form a surface.

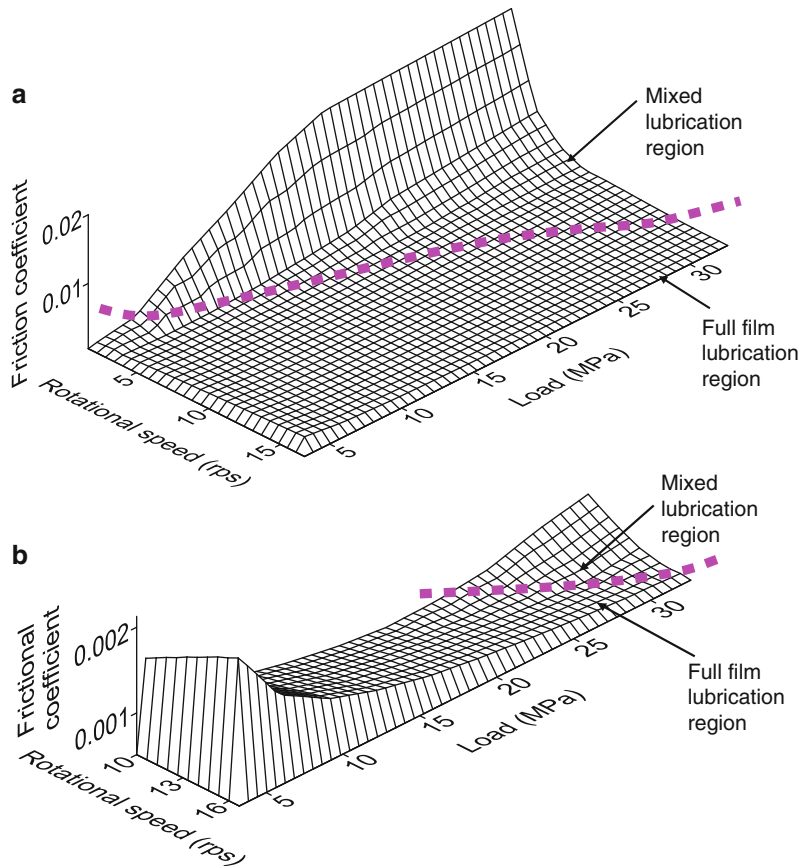
Effect of Load

A three-dimensional Stribeck surface can be constructed to show the relationship among friction coefficient, applied load, and rotational speed, as shown in Fig. 3, where the first axis is the load, the second axis is the rotational speed, and the third (vertical) axis is the friction coefficients in mixed and full-film lubrication regions under pairs of given load and rotational speed corresponding to Hersey numbers. The dotted line divides the Stribeck surface into two regions; the region above the curve is the mixed lubrication region, while the bottom region is in the full-film lubrication region. The shape of the surface at higher speeds is shown in an enlarged view.

Effect of Frictional Heat Transfer

Friction may cause temperature rise and alter film thickness due to the change in viscosity. Thermoelastic deformation should also be considered. In addition, in mixed lubrication, asperity contact occurs inevitably due to insufficiency of fluid film. All these can be included in the non-dimensional film thickness, defined as film thickness divided by root-mean-square roughness.

For mixed lubrication of a bearing with a free external surface, if the thermal deformation is caused by bearing



Stribeck Curves, Fig. 3 Stribeck surface showing the load, speed, and friction coefficient relationship (a) Rotational speed range 1.67 ~ 16.67 rps (b) Rotational speed range 10 ~ 16.67 rps (an enlarged view of (a))

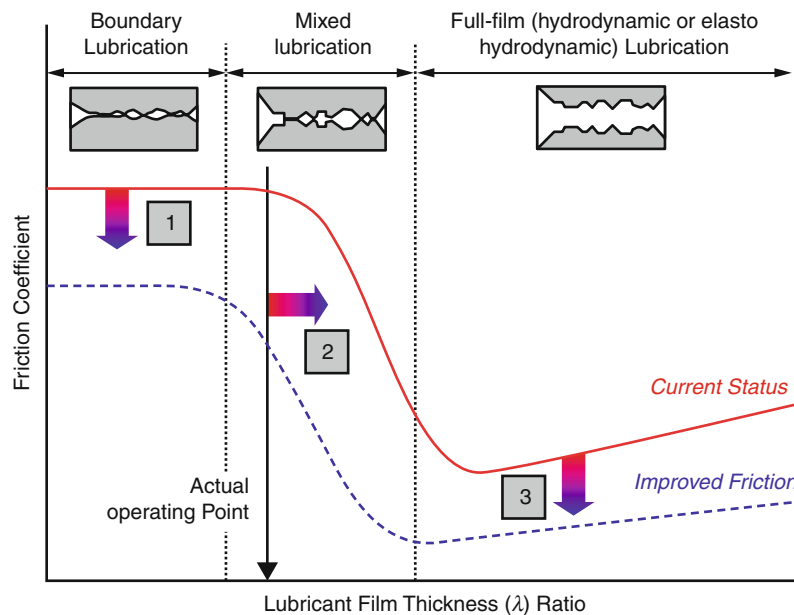
thermal expansion towards its external surface only, it produces a thicker film due to gap enlargement and the pressure-induced elastic deformation. In this case, if elastic and thermoelastic deformations are not considered, the film is the thinnest and the friction due to asperity sliding may be the highest, compared with the other cases. Elastic and thermoelastic deformations modify the film thickness, which in turn affects friction. Again, the case considering hydrodynamics only has the largest total friction coefficient because of its relatively thin film thickness and, therefore, the most severe asperity contact. Under the same operating conditions, the thermo-elasto-hydrodynamic (TEHD) case has the thickest fluid film and the lowest friction.

Key Applications

The concept of the Stribeck curve reveals the insight of friction in lubrication.

Division of Lubrication Regime

The Stribeck curve can be used for division of lubrication regimes (Luengo et al. 2002). Figure 1 shows a schematic Stribeck curve for a journal bearing system. It is interesting to see the existence of the minimum friction coefficient; the valley of the Stribeck curve marks the division between the mixed and full-film lubrications. The full-film lubrication friction is mainly due to viscous shear related to lubricant properties and surface topography that affects lubricant film thickness and resistance to hydrodynamic flows. The boundary lubrication regime is characterized by relatively high friction, which is primarily governed by the degree of surface interaction and the boundary layer formed through chemical and/or physical interactions among the lubricant molecules, additives, and surfaces. When the boundary film is removed, direct material contact may occur, which can result in high friction and, eventually, component or system failure. The mixed



Stribeck Curves, Fig. 4 Stribeck Curve as a function of non-dimensional lubricant film thickness in terms of film thickness over root-mean-square roughness. The Stribeck curve is a good indicator for the strategy of friction reduction and lubrication performance improvement (Martini et al. 2007)

lubrication regime is in between boundary and full-film lubrications, where friction decreases with the Hersey number until it reaches the minimum. This friction reduction is mainly due to the reduction of asperity contact, and at the frictional minimum, asperity contact diminishes.

Strategy of Friction Reduction

The Stribeck curve suggests three major ways to improve the lubrication performance and minimize friction between contacting surfaces in relative motion (Martini et al. 2007): (1) reducing boundary friction by improving boundary film performance by utilizing low-friction materials, surface coatings, and lubricant additives; (2) reducing the surface contact area by optimization of operating conditions and surface finish/texture; and (3) reducing hydrodynamic friction by improving lubricant rheological properties and surface topographic design. The effect of these friction reduction methods can be illustrated via the Stribeck curve as a function of the lubricant film thickness ratio, as shown in Fig. 4.

Determination of Liftoff Speed

Based on Khonsari and Booser (2010), a liftoff speed may be defined at the Hersey number corresponding to the minimum friction coefficient. Above this speed, asperity rubbing should have nearly no influence on friction.

This speed may be determined based on two trends of friction. Equation (3) may be simplified into the following expression using a friction partition parameter, ξ , varying between 1 and 0, where 0 is for full-film lubrication and 1 is for complete boundary lubrication with complete asperity contact:

$$f = -\xi f_c + (1 - \xi) f_{fl} \quad (4)$$

where f is the friction coefficient and subscripts c and fl are contact rubbing and fluid film, as indicated earlier. Here, the asperity rubbing friction, f_c , can be determined from a friction experiment, while the fluid-film friction, f_{fl} , can be estimated from (2) with considerations of fluid rheology and the influence of surface topography. The two terms of (4) can be plotted separately, and the intersection is roughly at the Hersey number corresponding to the liftoff speed (Khonsari and Booser 2010); however, note that f_c varies with film thickness and f_{fl} is related to the properties of the fluid, and film thickness and pressure gradient influenced by surface asperities.

Cross-References

- Average Reynolds Equation
- Elastohydrodynamic Lubrication (EHL)
- Flow Factors for Average Reynolds Equation

- [Mixed EHL](#)
- [Reynolds Equation](#)

References

- A. Akbarzadeh, M.M. Khonsari, Effect of surface pattern on stribeck curve. *Tribol. Lett.* **37**, 477–486 (2010)
- M.D. Hersey, The laws of lubrication of horizontal journal bearings. *J. Washington Acad. Sci.* **4**, 542–552 (1914)
- M.M. Khonsari, E.R. Booser, On the stribeck curve, in *Recent Development in Wear Presentation*, ed. by L. Friction, G. Nikas (Old City Publishing, Philadelphia, 2010), pp. 263–278
- X.B. Lu, M.M. Khonsari, The stribeck curve: experimental results and theoretical prediction. *J. Tribol.* **128**, 789–794 (2006)
- G. Luengo, J. Israelachvili, S. Granick, Generalized effects in confined fluids: new friction map for boundary lubrication. *Wear* **200**, 328–335 (2002)
- A. Martini, D. Zhu, Q. Wang, Friction reduction in mixed lubrication. *Tribol. Lett.* **28**, 139–147 (2007)
- N. Patir, H.S. Cheng, An average flow model for determine effects of three dimensional roughness on partial hydrodynamic lubrication. *ASME J. Lubr. Technol.* **100**, 12–17 (1978)
- J. Peklenik, New development in surface characterization and measurement by means of radon process analysis. *Proc. Insts. Mech. Eng.* **182**(3), 108 (1967)
- R. Stribeck, Die wesentlichen Eigenschaften der Gleit und Rollenlager. *Zeitschrift des Vereines Deutscher Ingenieure*, **36**, pp. 1341–1348; **46**, pp. 1432–1438; **46**, pp. 1463–1470 (1902)
- Y. Wang, Q. Wang, C. Lin, F. Shi, Development of a set of stribeck curves for conformal contacts of rough surfaces. *Tribol. Trans.* **49**, 526–535 (2006)

Studies on Friction, Wear, and Lubrication

- [Tribology](#)

Stylus Profilometry

ROLF KRÜGER-SEHM

Physikalisch-Technische Bundesanstalt Braunschweig,
Working Group 5.14, Roughness Measurement Methods,
Braunschweig, Germany

Synonyms

[Contact stylus method](#)

Definition

Contact stylus profilometry is a surface topography measurement method where a mechanically contacting

element scans over the surface boundary of a solid state body. The place of contacting and its height in a reference coordinate system is monitored in a regular raster.

Scientific Fundamentals

Measuring Principle

Comprehensive description are given in VDI Guideline 2602-2 (VDI/VDE-Society Measurement and Automation 2008) and by Volk (2005). Particular descriptions can be found in Whitehouse (1994, 2002) and in Blunt and Jiang (2003).

According to the measuring principle shown in Fig. 1, a stylus touches with its tip the surface to be measured. When the surface is displaced relative to the stylus in a horizontal direction, the stylus follows the height changes of the surface in the vertical direction. The height displacement measuring sensor converts the vertical position into a signal, which can be processed further. During the horizontal scanning, the position and height are sampled simultaneously in equidistant time or space intervals. The result is a profile or area data set, which is stored in the memory of the controlling computer. In the further processing of the data, in a first step some instrument corrections are applied, which are briefly described below under “[Deviations in Real Instruments](#).” For the determination of surface topography parameters, the set of data (or a subset of it) is evaluated by following agreed calculation rules in ISO 3274 (ISO 3274 1997).

Hardware Components of Contact Stylus Instruments

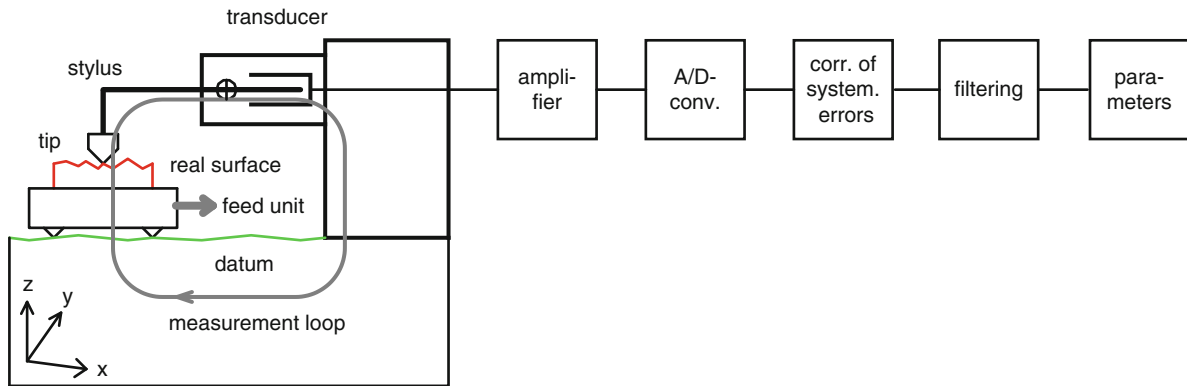
Stylus

The stylus tip is the most critical element in the contacting process and, therefore, needs special care during its design and use. It must be well defined in its geometry, mechanically stable, and wear resistant. Its design (cf. Fig. 2) follows accepted rules, which take into account different, partially contradicting aspects like the ability to detect fine details, mechanical stability, averaging behavior, manufacturing costs, and ease of use. Rules for its selection are given in international standards like ISO 3274 or ISO 25178-601 (ISO 3274 1996; ISO 25178-601 2010).

In contact stylus instruments, the usual rage of values for the tip radius is between 0.1 and 50 μm , the cone angle is preferably 60° or 90°. The material of the tip is usually diamond, sometimes sapphire.

Stylus Guiding Mechanism

The movement of the stylus follows the height change of the surface but is also forced by the stylus guiding



Stylus Profilometry, Fig. 1 Measuring principle of a surface contact stylus instrument

mechanism. In Fig. 3, some of the most typical principles are illustrated (not to scale).

These guiding mechanisms are different in their mechanical complexity and need more or less correction effort resulting from the deviation of the straightness of movement. They are connected with different amounts of moving mass of the guiding mechanism and the carrying capability for the mass of the stylus and the measuring sensor components.

Type	Advantage	Disadvantage	Remarks
Cantilever	Most freedom to observe contacting point	Arcuate movement correction, preferred scanning direction	Most common
Parallel spring	Movement with little straightness deviation	Contacting point can only be observed from side	
Double parallel spring	No straightness deviation	Increased moving mass	
Circular air bearing	No straightness deviation	Additional carrying facility	Independent from scanning direction

Height Displacement Measuring Sensor

The guiding mechanism is correlated with the design of the sensor, which converts the displacement of the stylus into a signal. Some examples of the most usual principles are shown as not to scale symbols in Fig. 4.

They are based on inductive differential inductive principles (Figs. 4a, b), differential capacitive transducers (Fig. 4c), or grating interferometer (Fig. 4d). The sensors measure the displacement of the stylus directly or via levers that increase or reduce the displacement. Aspects for their selection are, e.g., resolution, linearity, measuring range, dynamic range, or immunity against external electromagnetic or optical distortions.

Scanning Driving Mechanisms

For the relative movement between surface and stylus either the specimen or the contacting system must be moved in one lateral direction. The driving mechanism has to introduce a force for the movement following the datum without influencing the other spatial directions, especially not the height measurement.

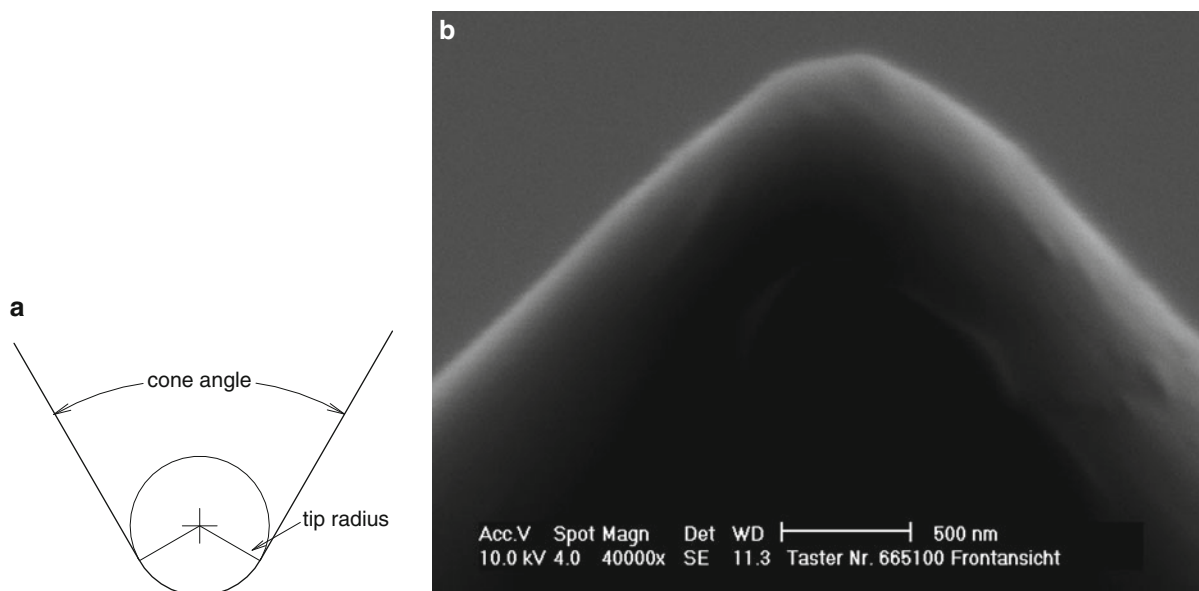
Typical driving mechanisms are lead screw, linear motor, voice coil, hydraulics, and piezo motors.

The datum for one axis contact stylus instrument is a straight line. To maintain the nearly ideal mechanical realization, e.g., of a lineal, it is necessary to decouple it from the driving mechanism with its usually inferior mechanical specifications. The decoupling mechanism determines the accuracy of the horizontal positioning of the stylus tip.

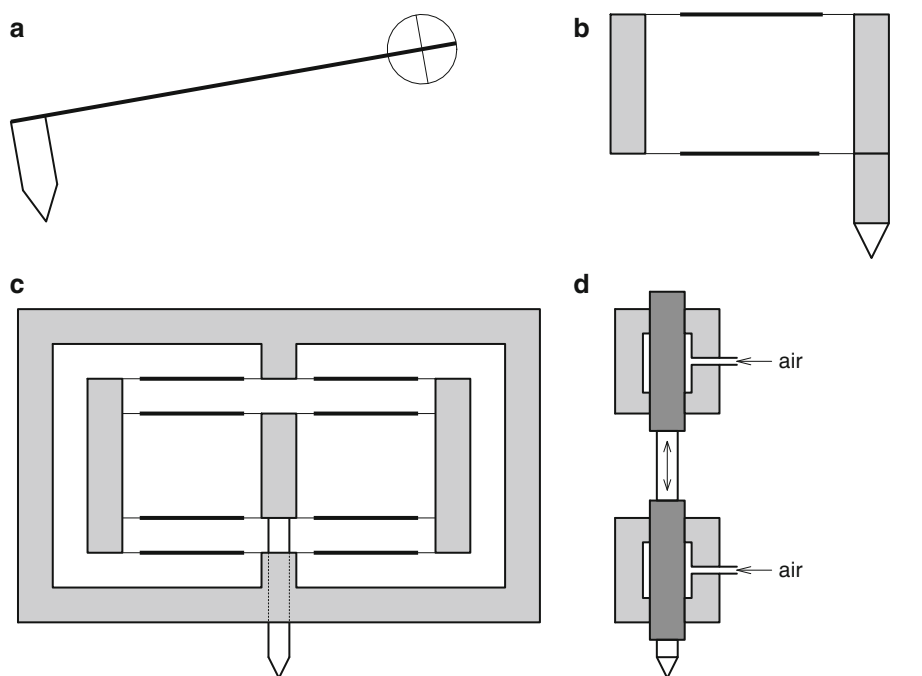
Arrangement of Displacement Measurement

The position of the lateral displacement measuring facility is ideally in line with the contacting point in the moving direction following Abbé's principle. In practice there are more or less deviations from this ideal.

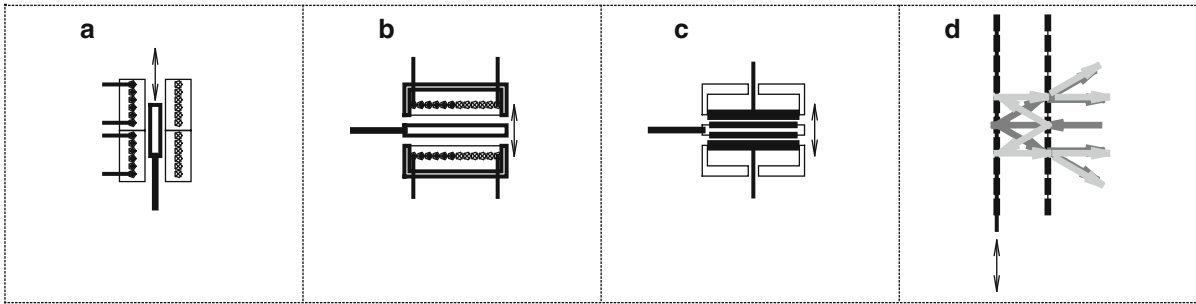
The simplest method is to have no measuring sensor and to trust in speed constancy and a stable sampling time interval.



Stylus Profilometry, Fig. 2 (a) Geometry of stylus tip, design. (b) SEM picture



Stylus Profilometry, Fig. 3 Typical stylus guiding mechanisms; (a) cantilever type (b) parallel spring (c) double parallel spring (d) circular air bearing



Stylus Profilometry, Fig. 4 Basic principles of displacement sensors; (a) linear differential inductive, (b) differential eddy current, (c) differential capacitor, (d) grating interferometer

Commonly used scanning displacement measuring methods include measuring the rotation of the spindle or the position of the driven part (nut) or the moving table itself by a linear line scale. For metrological purposes a laser interferometer directly measures the displacement of the table.

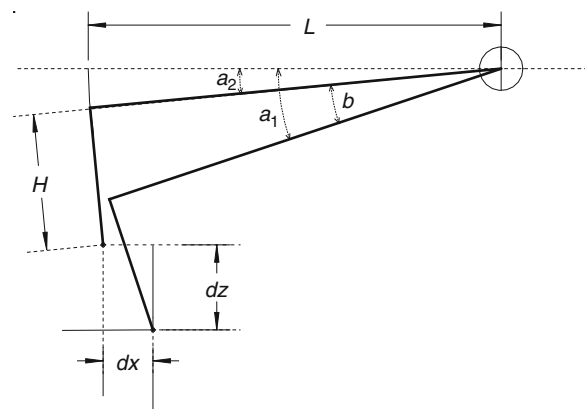
Two Axes Profiling

For the measurement of areas, a two axes relative movement between tip and surface and a datum plane is necessary. A stacked xy-table, each with its own driving and measuring facility, is a typical solution. A linear displacement of the tip in one horizontal direction and movement of the specimen in the orthogonal horizontal direction is another rather simple solution, which allows a dedicated localization of error sources. In a one plane version, the table with the specimen is movable on a common plane on an air bearing or on a slide bearing. The displacement driving mechanisms, their measuring systems, and the guides can be stacked or individually refer to the common plane. So there exists a large variety of sophisticated realizations, which cannot be listed here in detail.

Deviations in Real Instruments

Long and Short Wavelength Deviations

In the measurement loop, going through the contacting point, there are influences that are not as ideally stable during measurement as described above. The origin is a movement of the base of the tip due to its displacement sensor, short or long wave deviations of the datum from the ideal geometry, distortion by the driving mechanism, or thermal effects. These all can cause additional erroneous signals. Dependent on their wavelength, these errors are interpreted as topography like roughness, straightness, or form deviations. By a series of calibration specimens

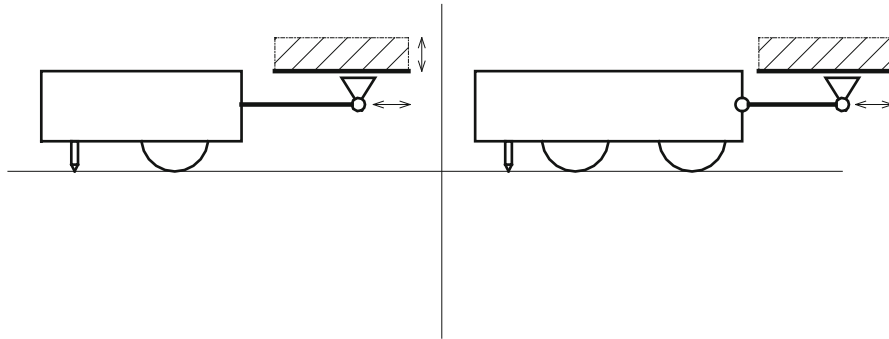


Stylus Profilometry, Fig. 5 Arcuate movement of cantilever type stylus tip; Cantilever length L , stylus height H , angular position α_1 , deviation of horizontal sampling position about dx by height change dz

(ISO 5436-1 2000) these influences can be identified in offline measurements and corrected for, if the influences are reproducible. A special case is noise in the roughness bandwidth range, which is statistically superimposed to the useful signal. Here compensation is not possible. This influence can only be reduced by design of the instrument or it must be taken into account as an influencing source in the calculation of uncertainty.

Influence of Stylus Tip Geometry

Because of the finite diameter of the stylus tip, the movement of an arbitrary point of the tip, e.g., the center of the tip sphere, cannot follow ideally all details of the surface topography. The movement of this representative point is the mechanical surface dilated by the tip sphere as structuring element (Srinivasan 1998). In the former standardization framework represented by, e.g., ISO 3274 (1996),



Stylus Profilometry, Fig. 6 Skidded contacting system

this profile changing was accepted as a nonlinear systematic error, which was common to all measurements that follow the standardized measuring and evaluation conditions. The Geometric Product Specification standards (GPS) will compensate for the dilatation by the tip radius by an erosion process on the contacted profile with a sphere having the nominal radius of the contacting stylus tip as structuring element. This correction has its limits in the minimum radius of concave surface curvatures becoming the radius of the structuring element.

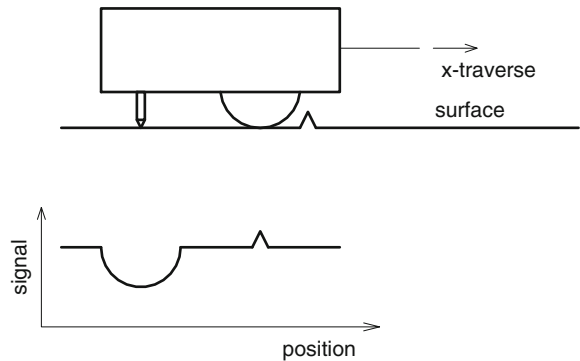
Contacting Force

The contacting of the surface is connected with a force, which is specified in ISO 3274 by about $0.7 \mu\text{N}$ in the nominal zero position of the height displacement measuring sensor. The stylus guiding mechanisms described above under “[Stylus Guiding Mechanism](#)” mostly acts as a spring, increasing or decreasing this force, when the stylus is moved out of this position. Even if the zero force is reduced by an additional compensating spring, the residual force has a spring elongation behavior connected with reduced moving range.

The contacting force has in practical use some consequences: Dependent on the hardness of the material, the surface is scratched (cf. [Fig. 5](#)). Therefore calibration standard specimens are manufactured out of hardened steel, glass, quartz, or hard metal.

When a step of dissimilar material is contacted, the parts of the step are compressed in different way, so that the measured step height differs dependent on the contacting force.

In modern contacting system design, the force is adjusted by static electromagnetic or electrostatic methods in conjunction with the AC displacement measuring technique.

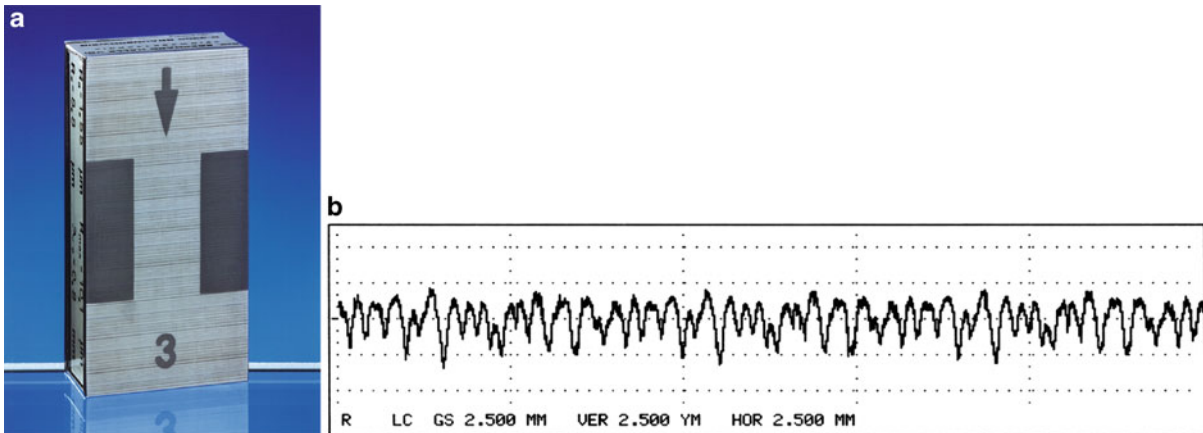


Stylus Profilometry, Fig. 7 Example for an error by single skid system with skid in feed direction

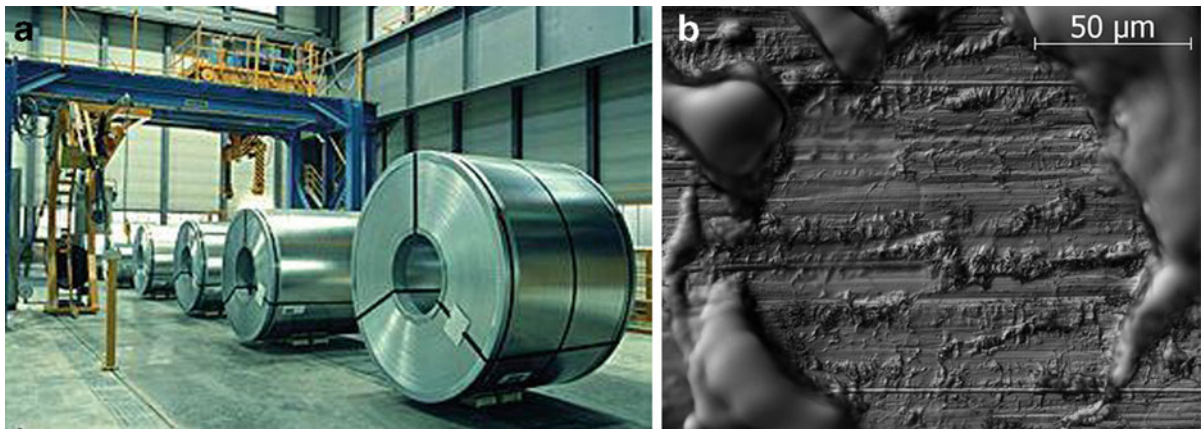
Arcuate Movement

In the most commonly used stylus guiding mechanism, the cantilever type, the stylus tip moves on an arc when it is forced to a vertical displacement (cf. [Fig. 5](#)).

The horizontal component dx is mostly negligible during classical roughness measurement with amplitudes up to $10 \mu\text{m}$, but becomes obvious when larger amplitudes are measured as realized in modern form measuring instruments with measurement ranges up to 10 mm . The so-called arc correction calculates the actual position of the stylus tip from the measured height value (representing the angular position of the stylus) and the geometry of the stylus. The result is an x, z -coordinate of the tip associated with the actual sampling position of the instrument. Thus, the sampling points representing the surface are no longer equidistant. For the calculation of surface texture parameters and the preceding wavelength filtering, a constant distance of the sampling point is



Stylus Profilometry, Fig. 8 (a) Calibration specimen (b) Measured profile. The systematic repetition of the profile is due to a measurement scheme well defined in calibration procedures



Stylus Profilometry, Fig. 9 (a) Coil of metal sheet and detail of surface (Source: Salzgitter Stahl AG, Germany) (b) Craters on the surface (manufactured by electron beam discharge), diameter of about 100 μm , wall heights of about 3 μm . A scratch of a measurement trace is visible by the applied optical contrast of Differential Interference Contrast (DIC). The depth of the trace is about 10 nm

assumed, and an interpolation into a new sampling point raster is applied.

Skidded Contacting System

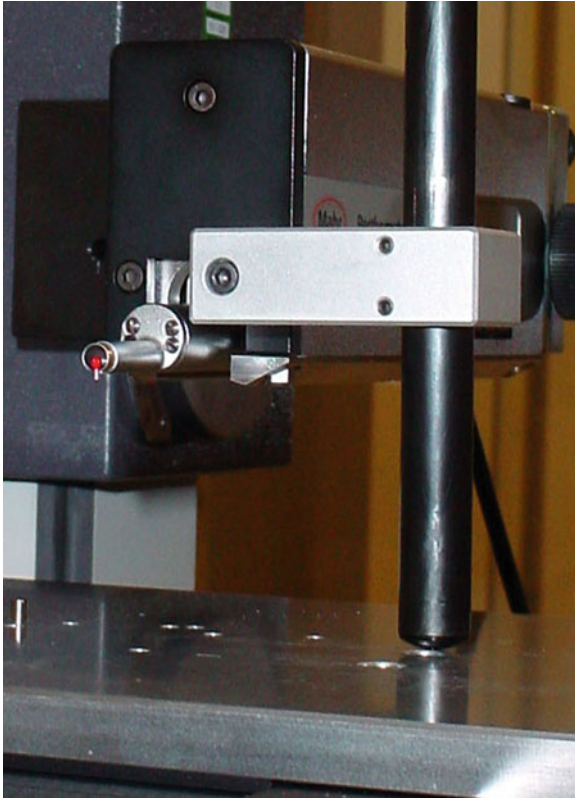
The measurement loop of the contact stylus instrument can pick up noise. In a skidded contacting system as illustrated in Fig. 6, the body of the contacting stylus system is placed on the surface via a skid. The feed unit draws the system over the surface to be measured.

The measuring stylus tip in the neighborhood of the skid contacts and measures the surface as described above. The surface itself, averaged by the skid, acts as the

reference. By this shortest possible measurement loop, this contacting system is extremely insensitive to mechanical distortions from the environment. However, this contacting system generates systematic errors dependent on the arrangement of the skid and stylus tip in relation to the scanning direction and the features on the surface. An example for this behavior is shown in Fig. 7.

Key Applications

Contact stylus instruments are used in a very wide field of measurement applications, covering engineering surfaces with roughness of about $R_z = 100 \mu\text{m}$ down to



Stylus Profilometry, Fig. 10 Damping facility in the measurement loop

single micrometers on the surface of injection needles of diesel engines. A calibration specimen contains a representative surface with an irregular structure of defined roughness. A view and a measured profile is shown in Fig. 8.

Another industrial application with a large commercial background is the measurement of metalsheet surfaces e.g., for cars and household utensils. The surface must have a dedicated structure to fulfil certain functional demands. An example of one of these demands is a bright appearance of the surface after paint coating. This must be assured between the commercial partners before coating by agreed contact stylus measurement on samples taken from the coils shown in Fig. 9a. A surface detail is shown in Fig. 9b.

In Fig. 10, a laboratory setup for a calibration measurement is shown. With intent to reduce the noise in the measurement loop a damping rod is mounted between the housing of the feed unit and the carrier of the specimen. The damping is realized by a loose coupling of the rod in its holder.

Cross-References

- [Accuracy of Surface Topography Characterization Tools](#)
- [Disk Roughness and Defect Monitoring](#)
- [Filtration of Surface Measurement Data](#)
- [Surface Roughness](#)

References

- L. Blunt, X.Q. Jiang (eds.), *Advanced Techniques for Assessment Surface Topography* (Kogan Page Science, London, UK, 2003). ISBN 1 9039 9611 2
- ISO 3274:1996, Geometrical product specification (GPS) – Surface texture: profile method – Nominal characteristics of contact (stylus) instruments 1996
- ISO 3274:1997, Geometrical product specification (GPS) – Surface texture: profile method – Terms, definitions and surface texture parameters 1998
- ISO 5436-1:2000, Geometrical product specification (GPS) – Surface texture: profile method – Measurement standards – Part 1: Material measures 2000
- ISO 25178-601:2010, Geometrical product specifications (GPS) — Surface texture: Areal — Part 601: Nominal characteristics of contact (stylus) instruments 2008
- V. Srinivasan, *Discrete Morphological Filters for Metrology*, in Proceedings of the 6th IMEKO ISMQC Symposium on Metrology for Quality Control in Production, TU Wien, Austria, Sept. 8–10, 1998
- VDI/VDE-Society Measurement and Automation, Roughness measurement using contact (stylus) instruments profile method; setup, measurement conditions, procedure, in *VDI Guideline 2602-2* (Beuth-Verlag, Berlin, 2008)
- R. Volk, *Rauheitsmessung, Theorie und Praxis* (Beuth-Verlag, Germany, 2005). 205 Seiten. ISBN 3-410-15918-5
- D.J. Whitehouse, *Handbook of Surface Metrology* (Institute of Physics, London, 1994). 998 pp
- D. Whitehouse, *Surfaces and Their Measurement* (Taylor and Francis, NewYork, 2002). ISBN 1-56032 969-6

Subdomain with Transformation Strain

- [Inclusions Subjected to Eigenstrain](#)

Subsurface-Originated Spalling

- [Probabilistic Life Prediction Models for Rolling Contact Fatigue](#)

SULF BT® Process (Similar Treatment)

► Sulfurizing at Room Temperature: The Aquasulf® Process

Sulfurizing at Room Temperature: The Aquasulf® Process

HERVÉ CHAVANNE^{1,2}, PHILIPPE MAURIN-PERRIER¹

¹HEF R&D, Z.I. Sud, Andrézieux-Bouthéon Cedex, France

²H.E.F. USA, Springfield, OH, USA

Synonyms

SULF BT® process (similar treatment)

Definition

AQUASULF® is an electrolytic sulfurizing process performed at room temperature in an aqueous solution. Parts are connected to the anode while cathodes are placed around it.

Scientific Fundamentals

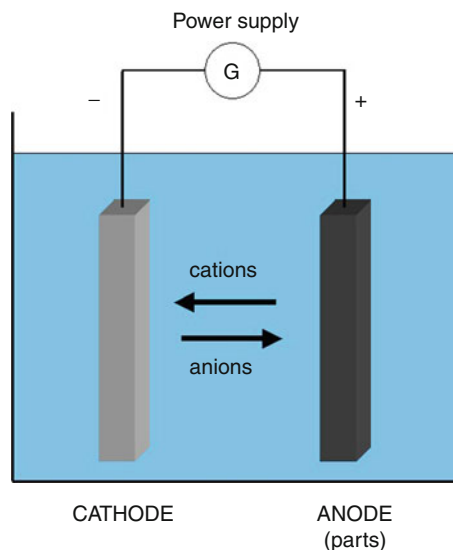
General Description

Sulfurizing of hardened steels is a well-known technique used to prevent seizure of mechanical components. The SULF BT® process has been used for more than 30 years on gears and differential components to reduce friction coefficient and seizure. Because of its relatively high cost, this process is limited to specific applications where price is not the main criteria or when other treatments such as phosphating cannot guarantee the desired friction properties.

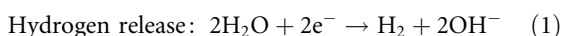
AQUASULF® is a more recent sulfurizing process, more affordable than SULF BT® but able to provide the same performance in use.

AQUASULF® Process

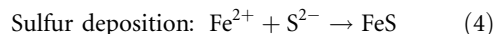
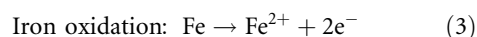
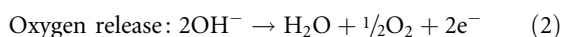
AQUASULF® is an anodic electrolytic sulfurizing process performed in aqueous solution at room temperature. Parts are connected to the anode (“+” side of the power supply), and cathodes in stainless steel are placed near the surfaces to be coated.



Reaction at the cathode:



Reactions at the anode:



The bath is a mixture of water, chloride salt, sulfide salt, and corrosion inhibitor. The total salts content shall not exceed 500 g/l. The power supply can provide either a pure DC or a pulsed DC current. The current density applied is in the range of 2–8 A/dm², depending on the shape and steel grade of the parts. The time of treatment is always less than 20 min.

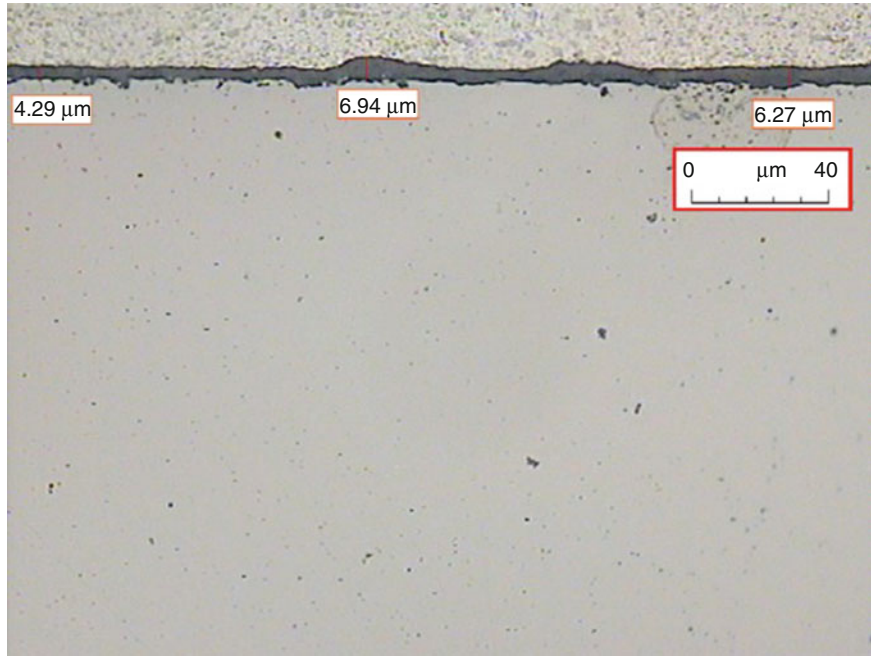
The sequence of the treatment is the following:

1. Degreasing in alkaline solution, connected to an oil separator
2. Acid pickling
3. Water rinsing
4. Sulfurizing at room temperature
5. Water rinsing
6. Temporary protection

Results

After the treatment, parts are coated with a black iron sulfide layer (FeS). The thickness of the layer is commonly between 3 and 10 µm (see Fig. 1).

While SULF BT® treatment induces negative dimensional changes (loss of ± 20 µm of the diameter),



Sulfurizing at Room Temperature: The Aquasulf® Process, Fig. 1 Cross section of a sulfurized low alloyed steel

AQUASULF® increases the diameter of the sulfurized parts in a range between 2 and 10 μm . The final roughness depends on the initial roughness, the base material, and the type of signal used (pure DC or pulsed DC). Final Ra parameter is between 0.5 and 1.5 μm .

Sulfurizing cannot be used to reduce problems of corrosion; the iron sulfide layer is very sensitive to oxidation and must be protected from moisture by an oily film.

Control of the Treatment

The Faville friction test is commonly used to control the efficiency of the treatment. A 6.5-mm diameter test piece rotates at 300 rpm between two jaws having 90°V notches into which the test piece fits. The two jaws squeeze the rotating part with a force that increases as the rotation proceeds. The test is considered complete when seizure occurs or when the measured load begins to fall, indicating creeping of the pin.

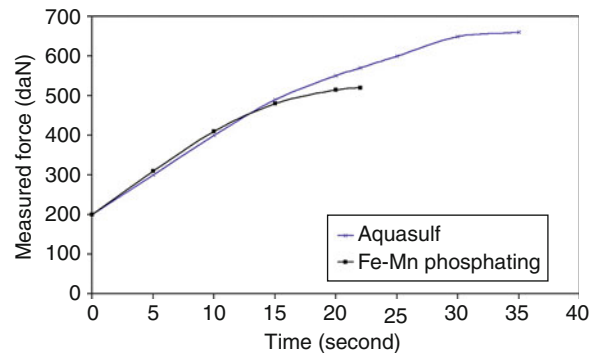
The Faville score (F) is calculated with the following equation:

$$F = \int_0^t (C_0 + Kt) dt \quad (5)$$

With

C_0 = initial load (200 daN)

K = load increase factor as function of time (for a standard tribometer, $K \approx 10 \text{ daN.s}^{-1}$)



Sulfurizing at Room Temperature: The Aquasulf® Process, Fig. 2 Examples of Faville test

In practice, the resultant force is measured every 5 or 10 s, and the following approximation is done:

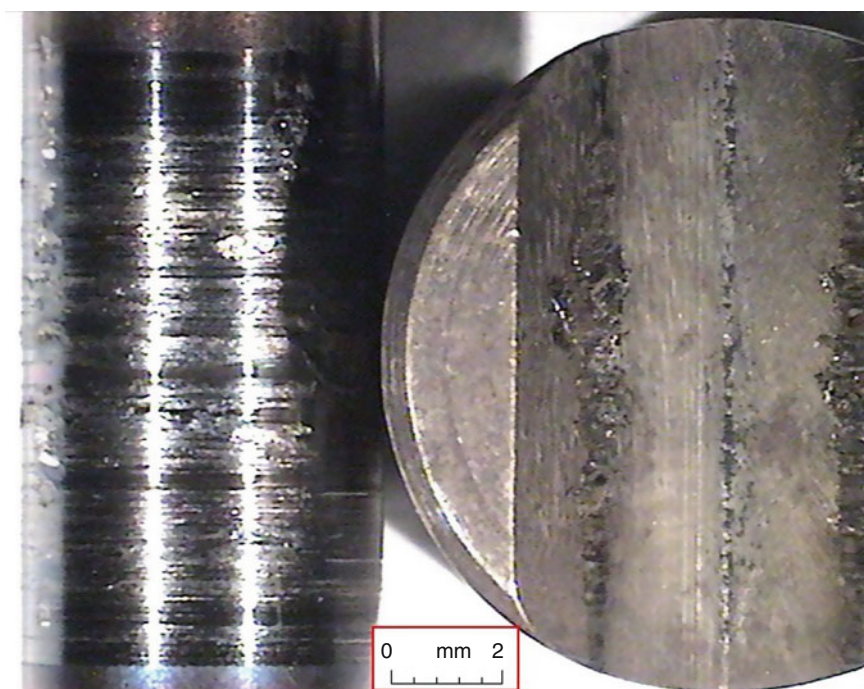
$$F = (C_n \Delta t \times \Delta t) + [CT \times (T - N\Delta t)] \quad (6)$$

With

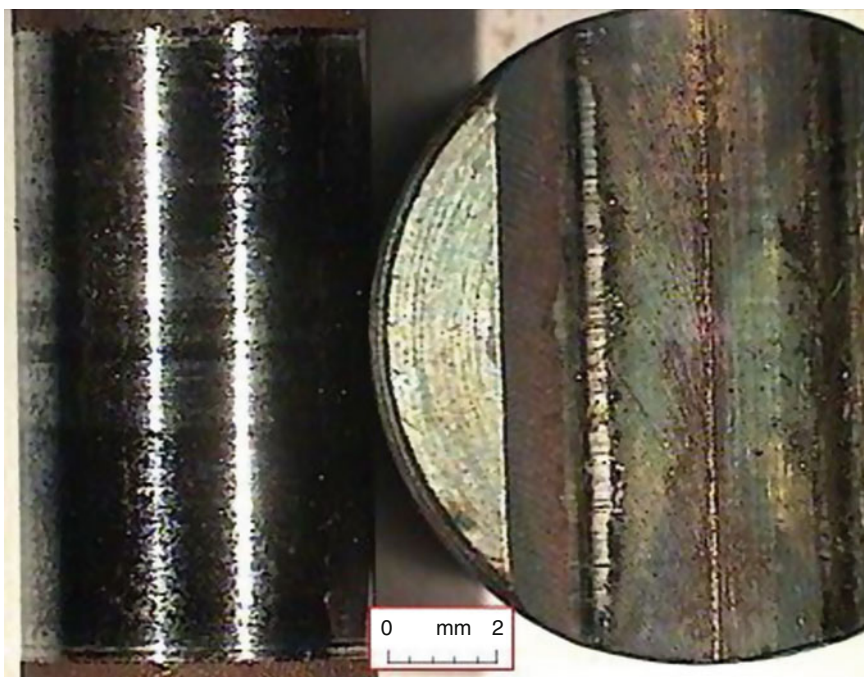
Δt = frequency of measurement of the resulting force (usually $\Delta t = 5$ or 10 s)

$C_{n\Delta t}$ = measured load at the time $n\Delta t$

T = end of test (seconds)



Sulfurizing at Room Temperature: The Aquasulf® Process, Fig. 3 Seizure on phosphated axis $F = 9,615 \text{ daN.s}$



Sulfurizing at Room Temperature: The Aquasulf® Process, Fig. 4 Creeping on sulfurized axis $F = 18,250 \text{ daN.s}$

C_T = final resulting force (before seizure or creeping)
 $N\Delta t$ = last entire period before the end of the test

The sulfurizing treatment is considered successful if the Faville test scores $F > 15,000$ daN.s. Figure 2 is an example of Faille's curves on an AQUASULF[®] sulfurized part and a Fe-Mn phosphated part with a corresponding score of, respectively, $F = 18,250$ daN.s and $F = 9,615$ daN.s.

After the test, seizure or creeping of the pin can be observed (Figs. 3 and 4).

Key Applications

The AQUASULF[®] process is often applied to gears where problems of wear, scuffing, and running noise are expected to appear, and also to the following:

- Differential components: pinion, planetary, cross head
- Transmission: thrust washers, spacer rings
- Universal joints

Materials suitable for treatment include case-hardened or induction-hardened steels, carburized steels, high-speed steels, cast irons, and sintered steels. To create the iron sulfide layer, the material must be a ferrous alloy.

Cross-References

- [Chemical Conversion Coatings](#)
- [Creep](#)
- [Friction Coefficient](#)

References

Science and Technology of Surface Coating, edition B.N. Chapman, J.C. Anderson, Sulf BT chapter, ISBN 0121683508

Sunflower Oil

- [Natural Oils as Lubricants](#)

Super D-Gun Coatings

- [Detonation Spray Coatings](#)

Superfinishing

- [Gear Surface Treatment](#)

Superhydrophobic Surfaces for Drag Reduction

RUI SHENG GUO, FENG ZHOU

State Key Laboratory of Solid Lubrication, Lanzhou
 Institute of Chemical Physics, Chinese Academy of
 Sciences, Lanzhou, People's Republic of China

Synonyms

[Low friction surfaces](#)

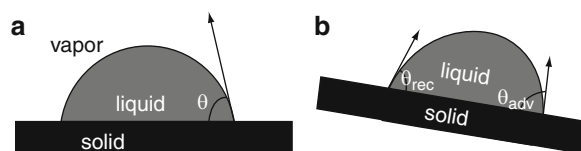
Definition

Superhydrophobic surfaces for drag reduction utilize a surface with superhydrophobic properties to reduce friction of a liquid flowing on it. Superhydrophobic surfaces are normally composite surfaces consisting of a large fraction of trapped air, thus generating boundary slippage and bringing about a shear-free air–water interface. Drag reduction significantly contributes energy saving and device efficiency in liquid transportation or other tribological systems.

Scientific Fundamentals

Basic Theory of Superhydrophobicity

Wettability is defined as the tendency of one fluid to spread on or adhere to a solid surface in the presence of other immiscible fluids. The common situation is that water displaces air on the solid surface. Wettability is characterized by the contact angle between the solid and a drop of liquid as shown in Fig. 1a. If the liquid wets the surface (referred to as the hydrophilic/oleophilic surface), the value of the contact angle is $0^\circ < \theta < 90^\circ$, whereas, if the liquid does not wet the surface (referred to as a hydrophobic/oleophobic surface), the value of the contact angle should be $90^\circ < \theta < 180^\circ$. The term hydrophobic/philic is usually used to describe the contact of a solid surface with water, while the term “oleophobic/philic” refers to wetting by oil

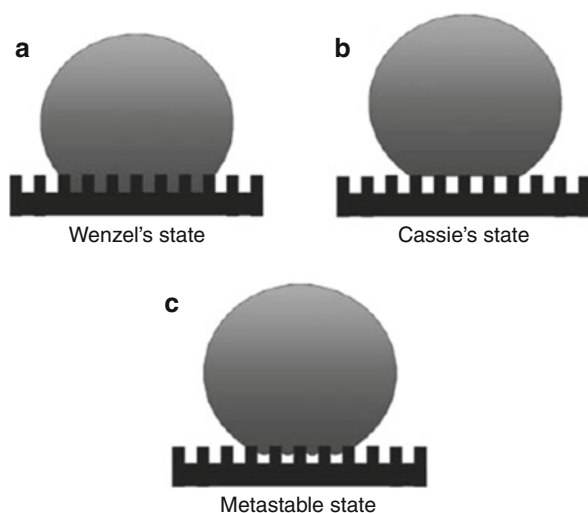


Superhydrophobic Surfaces for Drag Reduction, Fig. 1

(a) Contact angle between the solid and liquid surfaces,
 (b) contact angle hysteresis ($\theta_{adv} - \theta_{rec}$)

and organic liquids. A surface is considered superhydrophobic/superoleophobic if θ is greater than 150° and contact angle hysteresis is low. Contact angle hysteresis defines the difference between advancing and receding contact angles when a drop of liquid starts to move on a tilting surface as shown in Fig. 1b, and it reflects the adhesion behavior between liquid and solid (higher contact angle has stronger adhesion).

In fact, microscopically, there are different contact modes at the interface of liquid and solid. Figure 2a shows the Wenzel's state, where the droplet has intimate contact with the surface features. The adhesion on liquid/solid interface is obviously large. Figure 2b displays Cassie's state, where the droplet suspends on the top of the rough features with air trapped underneath; thus, the droplet is unstable and can roll off easily. However, in most cases, a water droplet may partially wet a surface and assume an intermediate state between Wenzel and Cassie states. Such an intermediate state of solid/liquid contact is referred as a metastable state (see Fig. 2c). This indicates that the external physical conditions can strongly affect the transition between the Cassie and Wenzel state, including the height of nanopillars, the spacing between pillars, and the intrinsic contact angle. Furthermore, transition between the two states could be achieved on the same microstructured surface when external stimuli (like a pressing force) exist.

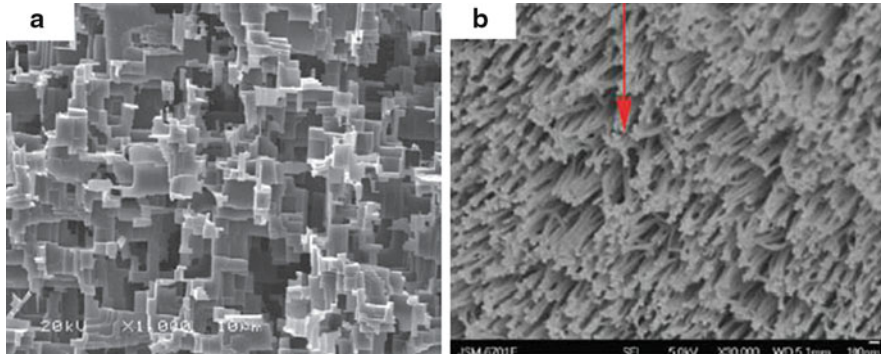


Superhydrophobic Surfaces for Drag Reduction, Fig. 2 Contact state of interface between liquid and solid (a) Wenzel's state, (b) Cassie's state, (c) Metastable state (Liu and Jiang 2010)

Fabrication of Superhydrophobic Surfaces

Superhydrophobicity has attracted a great deal of attention both in fundamental research and for potential applications including self-cleaning windows, windshields, exterior paints for buildings and navigation of ships, utensils, and antifouling. The focus here is on using a superhydrophobic surface for drag reduction.

Surface roughness plays a critical role in surface wettability. It influences the behaviors of liquid droplets both thermodynamically and hydrodynamically. Enhanced surface roughness will increase the surface's static contact angle and make the surface water repellent. However, it will sometimes increase the contact angle hysteresis and make it difficult for the droplet to roll off. Hence, it is very important to design and optimize surface roughness to obtain superhydrophobic surfaces with high contact angle and low contact angle hysteresis. To obtain the enhanced Cassie state superhydrophobicity, the study and simulation of biological objects with desired properties is referred to as "biomimetics." Biomimetics involves looking for engineering solutions from nature, mimicking them, and implementing them in practice. The lotus leaf is a typical example. The lotus emerges totally clean from muddy water. By observing the surface morphology, people found that the cooperation of the surface micro- and nanometer hierarchical structures and low-surface-energy hydrophobic wax-like material contributes to the superhydrophobicity. This finding is considered a great step forward in the field for the fabrication of artificial superhydrophobic surfaces. On the basis of our understanding of nature, a number of artificial hydrophobic surfaces have been fabricated with low surface energy materials and hierarchical structures using electrochemical methods, colloidal particles, photolithography, soft lithography, plasma treatment, self-assembly, and imprinting. As such an example, Wu et al. (2009) prepared a super-repellent surface by forming multiple-facet supported alumina nanowires with hierarchical micro/nanostructures (Fig. 3), which showed super-repellency towards a broad range of liquids after post-modification with perfluorosilane, including water, hexadecane, silicone oil, and crude oil. This provides a good example of biomimicking beyond nature since the superoleophobic surfaces are rarely found in nature. The complicated surface structure can be replicated into conventional polymeric coatings and materials (Liu et al. 2009), exhibiting superhydrophobicity and superoleophobicity as well after additional surface treatment. Inspired by nature, people have created a number of functional superhydrophobic materials. For example, anisotropic superhydrophobic surfaces are inspired by pigeon and goose feathers or



Superhydrophobic Surfaces for Drag Reduction, Fig. 3 Micro- and nanoscale hierarchical alumina **(a)** microscale multi-facet aluminum, **(b)** nanowire forests on a multi-facet mattress

bamboo leaves, superhydrophobic antifogging coatings are inspired by mosquito eyes; antireflective surfaces by moth eyes and cicada wings; and superhydrophobic surfaces that are highly adhesive mimic rose petals and gecko feet (Liu et al. 2010).

Fundamental of Flow

There are mainly two types of flow: laminar and turbulent. In a cylindrical conduit one can visualize the laminar flow as a series of co-axial cylinders oriented along the flow direction; such a flow structure is known as telescopic shear. The central part of the fluid has the highest velocity U . The velocity on the wall is zero, with the intermediate velocities in-between. A schematic representation is shown in Fig. 4a. Another way to create laminar flow is using two parallel plates with one moving and one stationary, as shown in Fig. 4b.

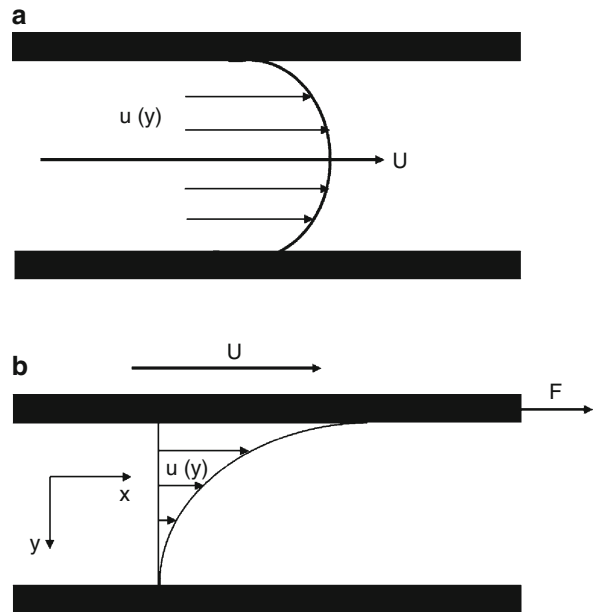
For the second mode, the fluid has the motion imposed by an applied shearing force F . When the top plate moves at the velocity U seen in Fig. 4b, velocities of fluid go down vertically along the y -axis from U (adjacent to the moving plate) to 0 (adjacent to the stationary plate). The following equation can be applied:

$$F = \mu \frac{U}{h} A, \quad (1)$$

where μ is the viscosity of fluid, A is the surface area to which the force is applied, h is the distance between parallel plates, and U/h is the vertical velocity gradient. The shear stress τ can be expressed,

$$\tau = \frac{F}{A} = \mu \frac{U}{h}, \quad (2)$$

Generically, when velocity distribution of $u(y)$ is at y site far from the stationary wall, the shear stress is



Superhydrophobic Surfaces for Drag Reduction, Fig. 4 **(a)** Laminar flow at velocity u in a cylindrical conduit, **(b)** laminar flow between parallel plates. The shearing force F acts on the top plate as indicated. The velocity u decreases going down along the vector y since the velocity at the bottom plate is necessarily 0 (Modified from (Brostow 2008))

$$\tau = \mu \frac{du}{dy}, \quad (3)$$

where du/dy is velocity gradient.

Reynolds number (Re) is normally used to discriminate between laminar flow and turbulent flow. For cylindrical conduit flow,

$$Re = \frac{Du_{av}\rho}{\mu}, \quad (4)$$

where D is the cylindrical conduit diameter, u_{av} is the average flow velocity, and ρ is the fluid mass density. For flow over a flat plate,

$$Re = \frac{Lu_{av}\rho}{\mu}, \quad (5)$$

where L is the length of flat plate. In general, natural transition occurs from laminar flow to turbulent flow regimes near a Reynolds number around 4,000 for cylindrical conduit flow and 500,000 for flow over a flat plate (Dean and Bhushan 2010). For values of Re much less than the above transition values (i.e., critical Reynolds number), flow is laminar. For larger Re values, the flow is turbulent.

Superhydrophobic Surfaces for Drag Reduction

The traditional view believes that no-slip boundary condition at the fluid/solid interface is an idealized paradigm, which assumes moderately strong attractive forces between the fluid and the wall. Thereby, the fluid atoms in bulk in attaining momentum and energy states differ from those of the solid boundary atoms in proximal contact. However, effects of surface tension, liquid evaporation, porosity, osmotic transport, van der Waals forces, and electrostatic forces may potentially result in true or apparent deviations from this classical picture. Even from a pure mathematical viewpoint, slip at the interface appears to be a more acceptable general notion than that of no-slip, since no-slip is a special case of slip with the magnitude of slip equal to zero! This fact was recognized by Navier more than a century ago, when he first

introduced the general notion of boundary slip by defining a slip length (L_S) as the distance behind the interface at which the fluid velocity extrapolates to zero (see Fig. 4). And the slip velocity, u_s , is proportional to the shear rate experienced by the fluid at the wall

$$u_s = L_S \frac{\partial u}{\partial y}. \quad (6)$$

In fact, the situations of $L_S = 0$ and $L_S = \infty$ are ideal cases, slip length is between 0 and ∞ in most situation (Fig. 5) (Chakraborty 2010).

The slip effect on the surface entails meaningful drag reduction for various flow conditions. Considering the situation of two parallel plates as shown in Fig. 4b, if one of the two surfaces has a slip length L_S , the drag reduction can be estimated as

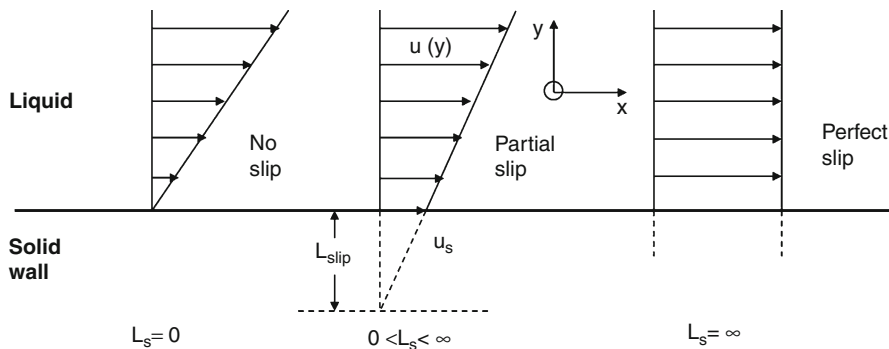
$$\frac{\tau_{slip}}{\tau_{no-slip}} = \frac{1}{1 + (L_S/h)}, \quad (7)$$

where τ_{slip} and $\tau_{no-slip}$ are the shear stresses at a wall when slip and no-slip boundary conditions are applied, respectively. In addition, according (7) the drag reduction can be calculated

$$D_R = \frac{\tau_{no-slip} - \tau_{slip}}{\tau_{no-slip}}. \quad (8)$$

Large drag reduction can be obtained as the gap between the plates becomes smaller, especially down to the range comparable to the slip length.

Usually, the rough structure of superhydrophobic surface is at the microscale. Due to the formation of space between solid posts, moderate roughness can lead to Cassie state contact between liquid and solid. It prevents the water from moving into the space, resulting in an air-water interface that is essentially close to shear-free.



Superhydrophobic Surfaces for Drag Reduction, Fig. 5 Concept of slip parameterized by the slip length, L_S (Modified from (Chakraborty 2010))

The resulting surface possesses a composite interface where momentum transfer with the wall occurs only at liquid–solid and not at liquid–vapor interfaces. Therefore, effective slippage will appear at the interface of liquid and solid. Simultaneously, experiments demonstrated that a vanishing slip length is found in the Wenzel state when the liquid impregnates the surface (Joseph et al. 2006). Therefore, superhydrophobic coatings with moderate patterning can result in an appreciable decrease of drag when liquid flows on the solid surface. Many experimental and numerical studies have reported that hydrophobic surfaces allow a noticeable slip ranging from nanometers to a micron in slip length and achieve drag reduction.

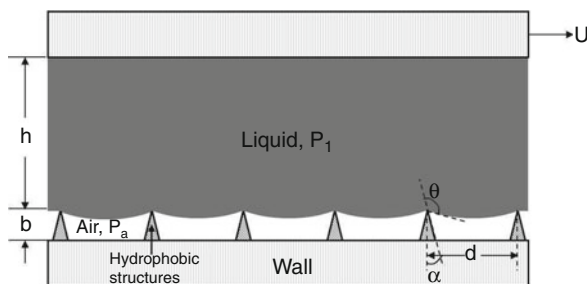
Application Study

Wetting-Related Drag Reduction

Choi et al. (Choi and Kim 2006) reported in detail about the design of superhydrophobic surface structures. They imagine Couette flow (fluid flows over a flat plate) in which an air layer separates liquid from a wall by the sharp tips of the hydrophobic posts, as shown in Fig. 6. Riding mainly over air, the liquid is expected to flow over the solid surface experiencing little friction. If one neglects the post structures (as an ideal case), a slip length L_S due to the pure air layer of thickness b , which is seen as the thickness of boundary layer, can be represented by

$$L_S = b(\mu_l/\mu_a - 1), \quad (9)$$

where μ_l and μ_a are the viscosities of liquid and air, respectively (Vinogradova 1995). A large effective slip is expected due to the sizable viscosity difference between liquid and air, larger with a thicker air layer. For example,



Superhydrophobic Surfaces for Drag Reduction, Fig. 6 Concept of large effective slip by a nanoengineered superhydrophobic surface in Couette flow. Liquid sits on hydrophobic structures by surface tension. The majority of the liquid boundary is with air, where shear stress is much smaller (Modified from (Choi and Kim 2006))

if the liquid is water and the air layer is 1 μm thick, the slip length would be 54 μm , disregarding the deviation from continuum at this scale.

For the design of the surface structures, consider conical posts of height b and cone angle α as shown in Fig. 6. The posts are assumed to form a square array with a pitch d . The meniscus is assumed to be of a spherical shape with contact angle θ (or advancing contact angle θ_A when the liquid pressure increases) on the side surfaces of the posts, balancing with the liquid pressure. The posts need to be tall enough so that the meniscus does not touch the bottom surface between posts. The posts also need to be populated densely enough, i.e., the pitch should be small enough so that the surface tension of the warped meniscus withstands the pressure in the liquid. If the pitch is low or loose, with increasing pressure, the state between liquid and solid will change from the Cassie state to the Wenzel state with liquid entering the space between pitches. So slip velocity reduces rapidly and loses the effect of drag reduction. By a simple geometrical calculation and the Laplace-Young equation, the post height b and the interpost pitch d to hold up the liquid meniscus against the pressure of liquid over air ($\Delta P = P_l - P_a$) can be obtained as

$$b > \frac{1 - \sin(\theta_A - \alpha)}{\sqrt{2}|\cos(\theta_A - \alpha)|} d, \quad d < 2\sqrt{2}\sigma \frac{|\cos(\theta_A - \alpha)|}{\Delta P}, \quad (10)$$

where σ is the surface tension of the liquid–air interface. For example, if the liquid is water ($\sigma = 0.0727 \text{ N/m}$ at 20°C), ΔP is 0.1 MPa ($\sim 1 \text{ atm}$), and $(\theta_A - \alpha)$ is 120° , the pitch d should be less than 1 μm , and the post height b should be larger than $\sim 0.2 \mu\text{m}$. It is further desired to make the posts sharp at the tip so that the liquid/solid contact area is minimized and slender in shape so that the contribution of air is maximized. Equation (10) serves as a key guideline in the design of proper geometry for the purpose at hand. According to this equation, the hydrophobic nanoturf surface was fabricated with 1–2 μm height and 0.5–1 μm pitch on a silicon wafer that was modified with Teflon by spin coating. Measured through a cone-and-plate rheometer system, the surface has demonstrated slip effects: a slip length of 20 μm for water and 50 μm for 30 wt.% glycerin liquid.

The above-discussed drag reduction is only limited to laminar flow. In turbulent flow, drag reduction has also been observed in experiments. Fundamentally, the effective reduction in solid–liquid boundary as a superhydrophobic drag reduction mechanism should be independent of whether the flow is laminar or

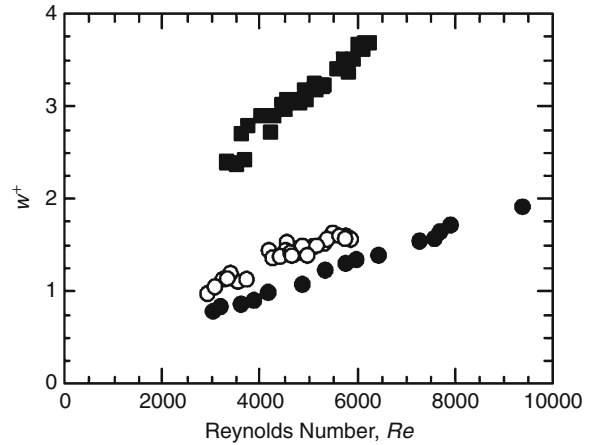
turbulent. In turbulent flows, a thin viscous-dominated sublayer exists very close to the wall. It extends to a height, measured in terms of wall units, viscous lengths, of

$$y^+ = y/v\sqrt{\tau_w/\rho} = 5. \quad (11)$$

where y is the height above the wall, v is the kinematic viscosity, τ_w is the wall shear stress, and ρ is the fluid density. In the viscous sublayer, the mean velocity increases linearly with position, $u^+ = y^+$. Changes in momentum transfer to the viscous sublayer can have a dramatic influence on the entire turbulent flow and can result in drag reduction for superhydrophobic surfaces.

Daniello et al. (2009) fabricated two types of superhydrophobic microridge geometries, which have been tested over a range of mean Reynolds numbers $2,000 < Re < 9,500$. Two geometries with 50% shear-free air-water interface coverage were considered. The first contains microridges $d = 30 \mu\text{m}$ wide and spaced $w = 30 \mu\text{m}$ apart (30–30) and the second contained microridges $d = 60 \mu\text{m}$ wide and spaced $w = 60 \mu\text{m}$ apart (60–60). Particle image velocimetry and pressure drop measurements were used to observe significant slip velocities, shear stress, and pressure drop reductions corresponding to drag reductions approaching 50%. At a certain Reynolds number, drag reduction increases with increasing feature size and spacing. And drag reduction promotes with further increasing Reynolds number.

Then they made further analysis for above results. Although the viscous sublayer thickness remains fixed in wall units, in dimensional form the thickness of the viscous sublayer decreases with increasing Reynolds number as $y_{vsl} = 5v\sqrt{\rho/\tau_w}$. Close to the wall, where viscous stresses dominate, the influence of the shear-free air-water interface extends to a distance roughly equal to the microridge spacing, w , into the flow. Thus, for the superhydrophobic surface to impact the turbulent flow, the microridge spacing must approach the thickness of the viscous sublayer, y_{vsl} , or in other words $w^+ = y^+ \approx 5$. As seen in Fig. 7, the microfeature spacing in wall units is at least $w^+ > 0.75$ for all the 30–30 surfaces and $w^+ > 2.4$ for the 60–60 surfaces. This means that the microfeature spacing is minimally 15–50% of viscous sublayer thickness almost immediately after the turbulent transition (critical Reynolds number is about 2,500). Hence for 30–30 and 60–60 ridges, drag reduction is noticed almost as soon as a turbulent flow develops. As the Reynolds number increases and the thickness of the viscous sublayer is further reduced, the presence of the superhydrophobic surface will more strongly influence the velocity profile



Superhydrophobic Surfaces for Drag Reduction, Fig. 7 The microridge spacing in wall units, w^+ , as a function of Reynolds number. The data are taken from PIV measurements from a channel with a single superhydrophobic surface of $w = 30 \mu\text{m}$ and $d = 30 \mu\text{m}$ microridges (●) and from pressure measurements for flow through a channel with two superhydrophobic walls containing $w = 30 \mu\text{m}$ and $d = 30 \mu\text{m}$ microridges (○) and $w = 60 \mu\text{m}$ and $d = 60 \mu\text{m}$ microridges (■). A spacing of $w^+ = 5$ corresponds to the thickness of the viscous sublayer (Daniello et al. 2009)

within the viscous sublayer and reduce the momentum transferred from the fluid to the wall and the vorticity of the fluid at the edge of the viscous sublayer. Turbulence intensity is thereby reduced, increasing the drag reduction. The 60–60 ridges have better drag reduction effect compared with 30–30 ridges due to their larger microfeature spacing. Of course, saturation of the turbulent drag reduction is likely in the limit of very large Reynolds numbers where the microridges are much larger than the viscous sublayer. In this limit, the drag reduction should approach a limit of $D_R = w/(d + w)$ as momentum is only transferred from the solid fraction of the superhydrophobic surface and the viscous sublayer is thin enough that the no-slip and shear-free portions of the surface can be considered independently. For the present shear-free area ratios, this limit would be 50%.

Truesdell et al. (2006) conducted laminar Couette flow measurements near a regularly textured superhydrophobic surface and drag reduction on the order of 20% was achieved. Watanabe et al. (1999) carried out experiments on highly water-repellant walls formed by the coating of a fluoroalkane-modified acrylic resin with added hydrophobic silica. The coating resulted in a hydrophobic surface crisscrossed by microcracks of

10–20 μm in width. Pressure drop and velocity profile measurements demonstrated drag reduction up to 18% and slip lengths up to 450 μm for flows at Reynolds numbers between 500 and 10,000. Chen et al. (2010) applied superhydrophobic coating on surfaces of steel pipes, a pipe flow system was established to measure the drag and to test the durability of the micro-structure of superhydrophobic coating at average speeds varying from 1 to 6 m/s (Reynolds numbers vary from about 38,000 to 230,000). These test speeds are more practical as far as industrial applications in the real world. It is quite interesting that the superhydrophobicity of the coating shows its due characteristic of drag reduction at a higher speed for which turbulent effects are supposed to be more significant. Nevertheless, such a good feature of drag reduction at a higher speed disappeared after about 30 min.

Perspectives

In engineering applications, drag reduction by a superhydrophobic coating has been an important topic in recent years, because they can be applied in microelectromechanical systems (MEMS) and tribology devices, such as particle transport by well-technology fluids in the oil industry. The development of techniques that produce significant drag reduction in turbulent flows can have a profound effect on a variety of existing technologies. The benefits of drag reduction range from a reduction in the pressure drop in pipe flows to an increase in fuel efficiency and speed of marine vessels.

However, much room is left for further improvement and future investigations of hydrophobicity. Firstly, slip-page mechanism is still under debate. Several reasons have been proposed for the slip over hydrophobic surfaces, including a molecular slip, a decrease in the viscosity of the boundary layer, the small dipole moment of a polar liquid, and a gas gap or nanobubbles at the liquid-surface interface. Here, the gas gap between the structures is used to explain the generated liquid slip. And slip length increases sharply with decreasing solid fraction and increasing effective contact angle. However, Voronov et al. (2008) demonstrated that, for hydrophobic surfaces, there is not necessarily a positive correlation between increased contact angle and slip length. So the theory of superhydrophobic drag reduction need to further study. Secondly, experimental techniques should be improved to capture the microscopic nature of slip more accurately (most importantly, a consensus of magnitude should be achieved). Because drag reduction is the ultimate goal of these studies, hydrophobic models in large size should be tested and experiments such as pressure drop in a pipe

should be systematically measured. Thirdly, present experiments are mainly taken by designing two-dimensional (2D) superhydrophobic surface, while three-dimensional (3D) surface structures are rarely studied. Finally, most experiments are carried out in a low-Reynolds number channel flow at present, while the range of Reynolds number is quite large in practice. In a word, additional theories and experiments are required to achieve significant drag reduction.

Cross-References

- [Liquid Contact Angle Measurement](#)
- [Lubrication Considering Boundary Slip](#)
- [Reynolds Equation](#)
- [Reynolds Number](#)
- [Shear Dependence of Viscosity](#)
- [Surface Free Energy](#)
- [Surface Roughness](#)

References

- W. Brostow, Drag reduction in flow: review of applications, mechanism and prediction. *J. Ind. Eng. Chem.* **14**, 409 (2008)
- S. Chakraborty, *Microfluidics and Microfabrication* (Springer Science + Business Media, New York, 2010)
- J.H. Chen, C.C. Tsai, Y.Z. Kehr, L. Horng, K. Chang, L. Kuo, An experimental study of drag reduction in a pipe with superhydrophobic coating at moderate reynolds numbers. *EPJ Web Conf.* **6**, 19005 (2010)
- C.H. Choi, C.J. Kim, Large slip of aqueous liquid flow over a nanoengineered superhydrophobic surface. *Phys. Rev. Lett.* **96**, 066001 (2006)
- R.J. Daniello, N.E. Waterhouse, J.P. Rothstein, Drag reduction in turbulent flows over superhydrophobic surfaces. *Phys. Fluids* **21**, 085103 (2009)
- B. Dean, B. Bhushan, Shark-skin surfaces for fluid-drag reduction in turbulent flow: a review. *Philos Trans. R. Soc. A* **368**, 4775 (2010)
- P. Joseph, C. Cottin-Bizonne, J.M. Benoit, C. Ybert, C. Journet, P. Tabeling, L. Bocquet, Slippage of water past superhydrophobic carbon nanotube forests in microchannels. *Phys. Rev. Lett.* **97**, 156104 (2006)
- M. Liu, L. Jiang, Switchable adhesion on liquid/solid interfaces. *Adv. Funct. Mater.* **20**, 3753 (2010)
- X. Liu, W. Wu, X. Wang, Z. Luo, Y. Liang, F. Zhou, A replication strategy for complex micro/nanostructures with superhydrophobicity and superoleophobicity and high contrast adhesion. *Soft Matter* **5**, 3097 (2009)
- K. Liu, X. Yao, L. Jiang, Recent developments in bio-inspired special wettability. *Chem. Soc. Rev.* **39**, 3240 (2010)
- R. Truesdell, A. Mammoli, P. Vorobieff, F. van Swol, C. Brinker, Drag reduction on a patterned superhydrophobic surface. *Phys. Rev. Lett.* **97**, 044504 (2006)
- O.I. Vinogradova, Drainage of a Thin Liquid Film Confined between Hydrophobic Surfaces. *Langmuir* **11**, 2213 (1995)
- R.S. Voronov, D.V. Papavassiliou, L.L. Lee, Review of fluid slip over superhydrophobic surfaces and its dependence on the contact angle. *Ind. Eng. Chem. Res.* **47**, 2455 (2008)

- K. Watanabe, Yanuar, H. Udagawa, Drag reduction of Newtonian fluid in a circular pipe with a highly water-repellent wall. *J. Fluid Mech.* **381**, 225 (1999)
- W. Wu, X. Wang, D. Wang, M. Chen, F. Zhou, W. Liu, Q. Xue, Alumina nanowire forests via unconventional anodization and super-repellency plus low adhesion to diverse liquids. *Chem. Commun.* 1043 (2009)

Superlaminar Flow

► Turbulence Effect on Fluid-Film Bearing Lubrication

Super-Long Life Fatigue

WANG QINGYUAN, HUANG ZHIYONG

Department of Civil Engineering and Mechanics, Sichuan University, Chengdu, People's Republic of China

Synonyms

Gigacycle fatigue (GCF); High cycle fatigue (HCF); Super-long life fatigue (SLLF); Ultra-high cycle fatigue (UHCF); Very high cycle fatigue (VHCF)

Definition

Super-long life fatigue has assumed greater significance in recent years, particularly after it was established that a fatigue limit does not exist in many cases. In many industries, the required design lifetime of some components often exceeds 10^8 cycles. This requirement is applicable to aircraft, automobiles, railways, offshore structures, microelectromechanical systems (MEMS), and biomedical parts, which may experience nominal vibratory stress conditions over a long period of time, running up to several hundred million cycles. This is referred to as *super-long life fatigue*.

Scientific Fundamentals

Introduction

Super-long life fatigue (SLLF), also known as gigacycle fatigue (GCF) (Wang et al. 1999), very high cycle fatigue (VHCF), and ultra-high cycle fatigue (UHCF), is exemplified by the cycles that a Japanese Shinkansen (Bullet Train) experiences during 10 years of use (Murakami et al. 2000).

In the research of Murakami et al. (2000), Ebara and Yamada (1987), and Kikukawa et al. (1965), the metal alloys can fail after 10^7 cycles. The phenomenon of GCF failure was observed and the conclusion has been made that the S-N curve from 10^6 to 10^9 is not asymptotic (Bathias and Paris 2004). Therefore, the application of fatigue study beyond 10^7 cycles benefits fatigue endurance design for some key components of, for example, automobile engines, high-speed trains, and aircraft engines. Investigation of crack threshold and its propagation speed provides better understanding of crack growth behavior and more precise fatigue life prediction under SLLF load. The crack initiation mechanism in a SLLF regime can be clarified by SEM observation of the fracture surface, which informs engineering improvements in the metallic material.

It is common that turbine blades experience more than 10^{10} stress cycles by vibrating during their service life, and a car wheel, for example, is exposed to about 3×10^8 cycles during a lifetime of 300,000 km. The need for material conservation, for both economic and environmental reasons, requires that the service lives of applications be extended to as many cycles as possible. Determination of long life fatigue behavior is important for better understanding and design of components and structures. There is a growing interest in investigating very high cycle fatigue behavior (10^7 – 10^{10} cycles) and very low crack growth rates (10^{-12} – 10^{-9} m/cycle) in various materials (Wang et al. 1999, 2002; Murakami et al. 2000; Ebara and Yamada 1987; Kikukawa et al. 1965; Bathias and Paris 2004; Sakai et al. 1999).

Super-long life fatigue has assumed greater significance in recent years, particularly after it was established that a fatigue limit does not exist in many cases. However, the current fatigue code design curves were developed on the basis of data for fatigue life of less than 10^6 – 10^7 cycles, far below the 10^9 – 10^{12} cycles experienced in the industries that require super-long service life. Most engineering designs have hitherto been based on the assumption that ferrous materials exhibit a fatigue limit and for any cyclic stress below this value, they are unaffected and endure infinite life. This was usually believed to be in the vicinity of 10^6 cycles, where the S-N curve becomes nearly horizontal, primarily because time and cost constraints had ruled out conventional fatigue testing beyond 10^6 – 10^7 cycles. More recent studies (Wang et al. 1999, 2002, 2003; Murakami et al. 2000; Ebara and Yamada 1987; Kikukawa et al. 1965; Bathias and Paris 2004; Sakai et al. 1999; Prasannavenkatesan et al. 2009; Xue et al. 2007; Shiozawa et al. 2009), however, point to the fact that most materials including ferrous alloys

experience failure up to 10^9 cycles and above and that this phenomenon of super-long life fatigue is a result of sub-surface or internal crack initiation at the defect sites in materials. In other words, the concept of the so-called fatigue limit has all along been a myth and it would be wise to conduct high cycle fatigue testing up to gigacycle regimes. Further, it would henceforth be more appropriate to state the fatigue strength of a material for a given number of cycles rather than as an endurance limit.

The common form of presentation of fatigue data uses the S-N curve, where the total cyclic stress (S) is plotted against the number of cycles to failure (N) in logarithmic scale. Different forms of S-N curve in super-long life regime are shown in Fig. 1, including step-wise S-N shape for materials like high-strength steels, continued degradation shape for Al-alloys, and S-N curve with fatigue limit for some low-carbon steels.

Test Methods for Super-Long Life Fatigue

Although a good deal of high-cycle fatigue (HCF) data have been published in the form of S-N curves, the data in the literature have been limited to fatigue lives up to 10^7 cycles. Conventional electrohydraulic pistons cannot provide fatigue loads to specimens at frequencies above approximately 100 Hz. Time and cost constraints rule out the use of conventional fatigue tests of more than 10^7 cycles to check structural materials. A possibility of accelerated testing of specimens is considered by using high-frequency cyclic loading.

Development of higher-frequency testing machines proceeded slowly over the years, starting early in twentieth century (Roth and Roth 1987). Prior to 1911, the highest fatigue testing frequency was on the order of 33 Hz. Electrodynamic resonance systems appeared in 1911, when Hopkinson introduced a machine capable of 116 Hz.



Super-Long Life Fatigue, Fig. 1 Super-long life fatigue testing machines

In 1925, Jenkin tested wires of copper, iron, and steel at 2 kHz, using similar techniques. In 1950, Mason developed ultrasonic machine capable of 20 kHz, using magnetostrictive and piezoelectric transducers to fatigue testing. Until recently, the majority of ultrasonic fatigue tests operated at frequencies of approximately 20 kHz.

The ultrasonic fatigue technique provides a practical means of generating super-long life fatigue data. A 100–1,000 times testing time compression can be obtained with ultrasonic fatigue. For example, a gigacycle (10^9) test would require more than 1.6 years with a traditional 20 Hz servohydraulic testing machine. The same test takes only 14 h with an ultrasonic fatigue machine of 20 kHz. The time required to complete fatigue tests at different frequencies is shown in Table 1.

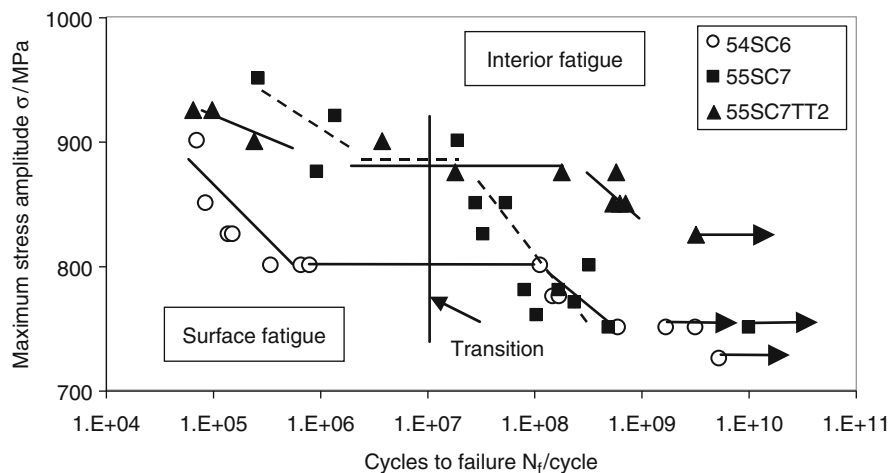
Ultrasonic fatigue is a resonant test method in which a large amplitude displacement wave must be established in a resonant specimen. The resonance system consists of a power generator chosen to provide a sinusoidal high-frequency signal (20 kHz) and the vibration energy, an ultrasonic converter to change the electronic signal into

mechanical vibration, and one ($R=-1$) or two acoustic horns to raise the vibration amplitude in the specimen to the level required for fatigue. To achieve resonance during testing, each part of the system was designed to resonate longitudinally at 20 kHz. The design of horns was developed with the need of different applications, including high/low amplification and sealing the entrance of an environmental chamber (Bathias and Paris 2004).

Super-long life fatigue is practically realized by the ultrasonic fatigue testing procedure. Conventional testing (on servo-hydraulic machines, rotating bending machines, etc.) operating up to 100 Hz is also widely used, even though it requires a long testing time. The maximum loading frequencies of conventional fatigue testing equipment are typically 100 Hz; with a closed-loop servohydraulic-type testing machine 1 KHz could be realized (Morgan and Milligan 1997). A magnetostrictive loading machine can reach 1.5 kHz (Davidson 1999). Various ultrasonic fatigue testing machines capable of 20–30 kHz frequency were developed by Bathias and Paris (2004) (Fig. 2a) and Stanzl. The USF-2000 ultrasonic

Super-Long Life Fatigue, Table 1 Testing time vs number of cycles at different frequencies

	Conventional fatigue			Ultrasonic fatigue
	1 Hz	10 Hz	100 Hz	20 kHz
10^7	4 months	12 days	30 h	9 min
10^8	3.2 years	4 months	12 days	1.4 h
10^9	32 years	3.2 years	4 months	14 h
10^{10}	320 years	32 years	3.2 years	6 days



Super-Long Life Fatigue, Fig. 2 S-N characteristics of high-strength Cr-Si steels in ultrasonic fatigue (Wang et al. 2002)

fatigue testing system (Fig. 2b) produced by Shimadzu is capable of testing to a frequency of 20 kHz and a stress ratio of $R = -1$. In Japan, super-long life fatigue tests were mostly performed using cantilever-type rotary bending fatigue testing machine (Fig. 2c) operating at a frequency of 52.5 Hz [39–40]. A newly developed multi-axial fatigue testing machine (PMF4-10, Fig. 2d) may operate simultaneous fatigue tests for four specimens at a frequency of 80 Hz.

Characterization of S-N Curve in Super-Long Life Fatigue

Generally speaking, the fatigue S-N curve for steels is always considered to be asymptotic (i.e., horizontal to the N axis); thus, no tests were carried out beyond 10^9 cycles in order to check the continued existence of this asymptote. Some materials display a fatigue limit or “endurance” limit at a high number of cycles (typically $>10^6$). Most other materials do not exhibit this response, instead displaying a continuously decreasing stress-life response, even at a great number of cycles (10^6 – 10^9), which is more correctly described by fatigue strength at a given number of cycles.

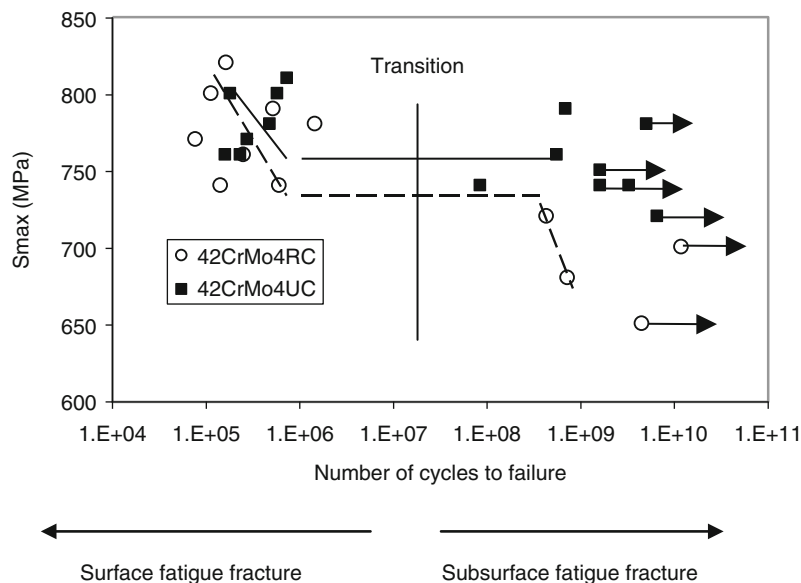
Figures 3–5 show some typical super-long life fatigue data between 10^5 and 10^{10} cycles obtained from an ultrasonic fatigue machine at 20 kHz and a rotating bending fatigue machine at 52.5 Hz, respectively. These typical S-N curves display a step-wise shape, generally with a second lower fatigue limit in the very high cycle regime ($>10^8$).

The most noticeable conclusion was that failure could occur beyond 10^7 , and even 10^8 stress cycles, and a fatigue limit could not be obtained until 10^9 cycles. It is therefore important to realize the risk of fatigue failure beyond 10^7 cycles. A significant change in the slope of the S-N curve was observed, accompanying the transition from surface to subsurface crack initiation.

Figure 6 shows the S-N data for 6061/T6, 7075/T6, and 2024/T3 aluminum alloys over a wide range, viz. between 10^4 and 10^9 cycles. The experimental results prove that fatigue failure in Al-alloys can very well occur up to 10^9 cycles, and in super-long life regime a decrease of fatigue strength with increased number of cycles still occurs in the Al-alloys, even if corrosion or temperature effects are excluded.

Mechanism of Super-Long Life Fatigue

It is well known that nonmetallic inclusions are usually present in all commercial steels as a result of the steel-making process. Inclusions, invariably being the sites for fatigue crack nucleation, play a crucial role in governing the fatigue lives of high-strength steels. While low-cycle fatigue of steels usually involves fatigue crack initiation from inclusions/defects on the specimen surface, in the high-cycle regime, the defects located within the specimen become favorable sites for crack initiation (Wang et al. 2002). The occurrence of internal cracking has been clearly observed in many materials containing inclusions/defects. Umezawa and Ishikawa (1994) and his group, in their



Super-Long Life Fatigue, Fig. 3 S-N characteristics of 42CrMo4 high-strength steels in ultrasonic fatigue (Sakai et al. 1999)

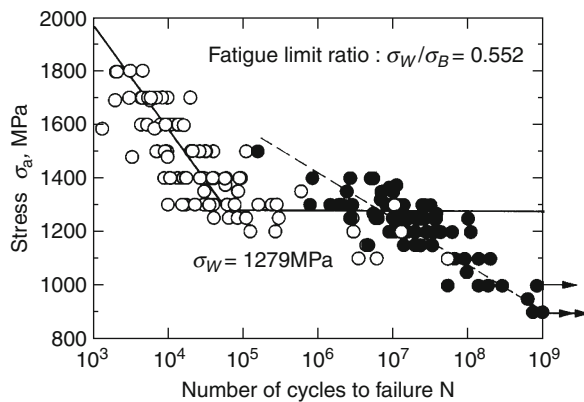
study on high-strength austenitic steels and titanium alloys at low temperatures, noticed subsurface crack initiation. Nishijima and Kanazawa (1999) obtained internal fatigue fracture in low alloy steel tested at 300–400°C up to 10^9 cycles. Subsurface crack initiation has been reported by Murakami et al. (2000) for Cr–Mo steel under tension–compression fatigue at 30–100 Hz, and by Nishijima and Kanazawa (1999) for a spring steel under rotary bending at 50 Hz. This phenomenon has also been observed in a variety of titanium and Ni alloys. However, the mechanism of fatigue cracking at the inclusion sites is still not fully understood. It is believed that fatigue fracture does not initiate from inclusions below a critical size. The idea of

the so-called *zero-inclusion steels* was proposed by Mitchell and Fukumoto (1991), where the inclusion size is well controlled to within 1 μm by adopting better steel-making practices. This seems to greatly improve the fatigue performance of such steels.

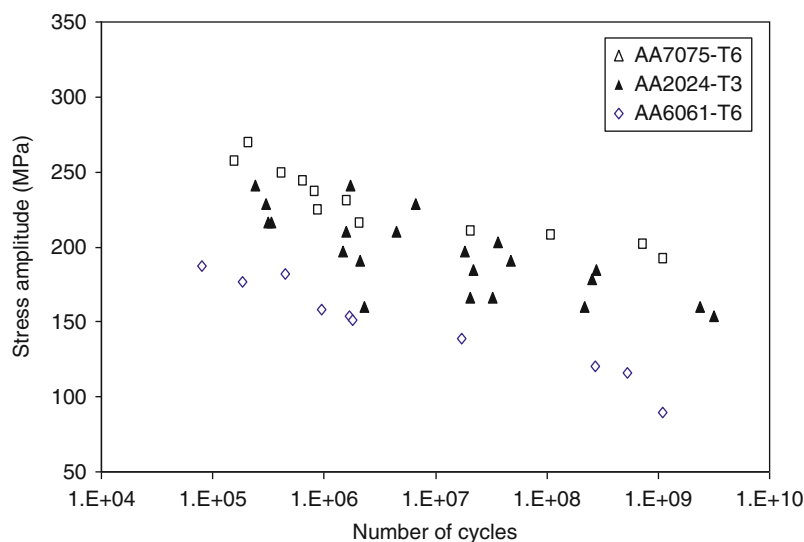
In super-long life regimes (beyond 10^7 cycles), the initiation sites were usually found at inclusions located in the interior of the specimens. On the other hand, fatigue initiation occurred on the surface. The stages of crack initiation, stable crack propagation, unstable crack propagation, and final failure are well defined, as shown in Fig. 7.

Some typical subsurface crack initiations in the 42CrMo4 low alloy steels and the Cr–V and Cr–Si spring steels are shown in Fig. 8. The stages of crack initiation, stable crack propagation, unstable crack propagation, and final failure are well defined. All of the subsurface crack initiation sites appear as flat, smooth features of facets. The fracture origin was identified by use of energy dispersive analysis. In the high-cycle regime ($>10^7$ cycles), all of the initiation sites were found at nonmetallic inclusions located in the interior of the specimen. The chemical composition of the inclusions was mostly sulfide. The inclusions range in size from 10 to 40 μm .

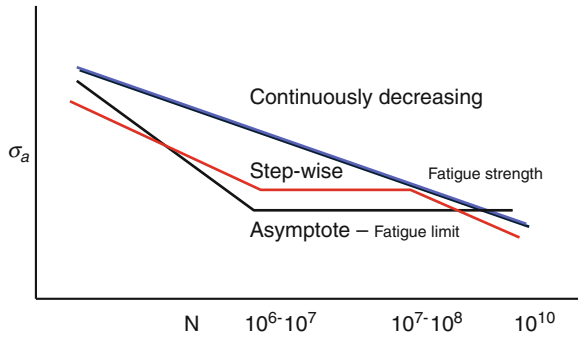
Super-long life fatigue fractures are generally rough in appearance and initiate predominantly from the surface in Al-alloys. Figure 9a shows an overall view of a typical high-cycle fatigue fracture surface of AA6061. A translamellar cleavage fracture mode was observed. The fracture surface also contained features of localized inhomogeneous planar slip deformation. The fracture was seen to initiate



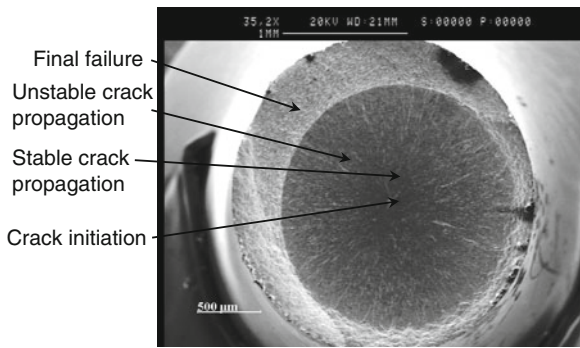
Super-Long Life Fatigue, Fig. 4 S–N characteristics of SUJ2 steel in rotating bending (Sakai et al. 1999)



Super-Long Life Fatigue, Fig. 5 S–N characteristics of Al-alloys in ultrasonic fatigue (Wang et al. 2006)



Super-Long Life Fatigue, Fig. 6 Different forms of S-N curve in super-long life regime



Super-Long Life Fatigue, Fig. 7 Overview of super-long life fatigue fracture surface (Kikukawa et al. 1965)

from micro-voids at the subsurface, leading to the formation of large micro-cleavage facets, comprised of a population of micro-voids of a range of sizes, during crack initiation and early growth (Fig. 9b, c). Coalescence of these expanding voids to form drawing lines in the cleavage facets and eventually macro-cracks was also observed. Fatigue failure thus seems to start with the formation, growth, and coalescence of interfacial voids, ending with the propagation of macro-cracks initiated at the base of the voids.

Again, the super-long life fracture surfaces of 7075/T6 alloy have a faceted appearance, although the facets are smaller than those in 6061/T6 alloy (Fig. 9), and significant interfacial void is observed in the fatigue crack growth process (Fig. 10a, b). Fatigue crack growth was initially transgranular, followed by a mixed transgranular-intergranular fracture. The mixed fracture surface exhibits cleavage flat areas with a population of micro-voids. The voids coalesce to form void lines randomly distributed through the alloy microstructure. The void lines can be considered as macroscopic cracks.

The super-long life fatigue fractures obtained in the cast magnesium alloys AE42 are shown in Fig. 11. The fractograph of the specimen tested to long life fatigue failure (2.1×10^8 cycles) is also shown. It is seen that fracture initiates slightly beneath the surface and is generally transcrystalline in nature. Cavities are also present.

Key Applications

Fatigue Life Prediction in Super-long Life Regimes

Because inclusions have a significant effect on the super-long life fatigue mechanism in high-strength steels, the need for a model is well understood. One such model to predict the fatigue life based on crack initiation as well as propagation is presented here.

Prediction of Fatigue Crack Initiation Life

A quantitative understanding of fatigue crack initiation (FCI) has been limited. Tanaka and Mura (1981) proposed a dislocation model for FCI from inclusions present in homogeneous materials. A simplified form of the model is as shown below. The cycles to crack initiation N_i is given by the expression:

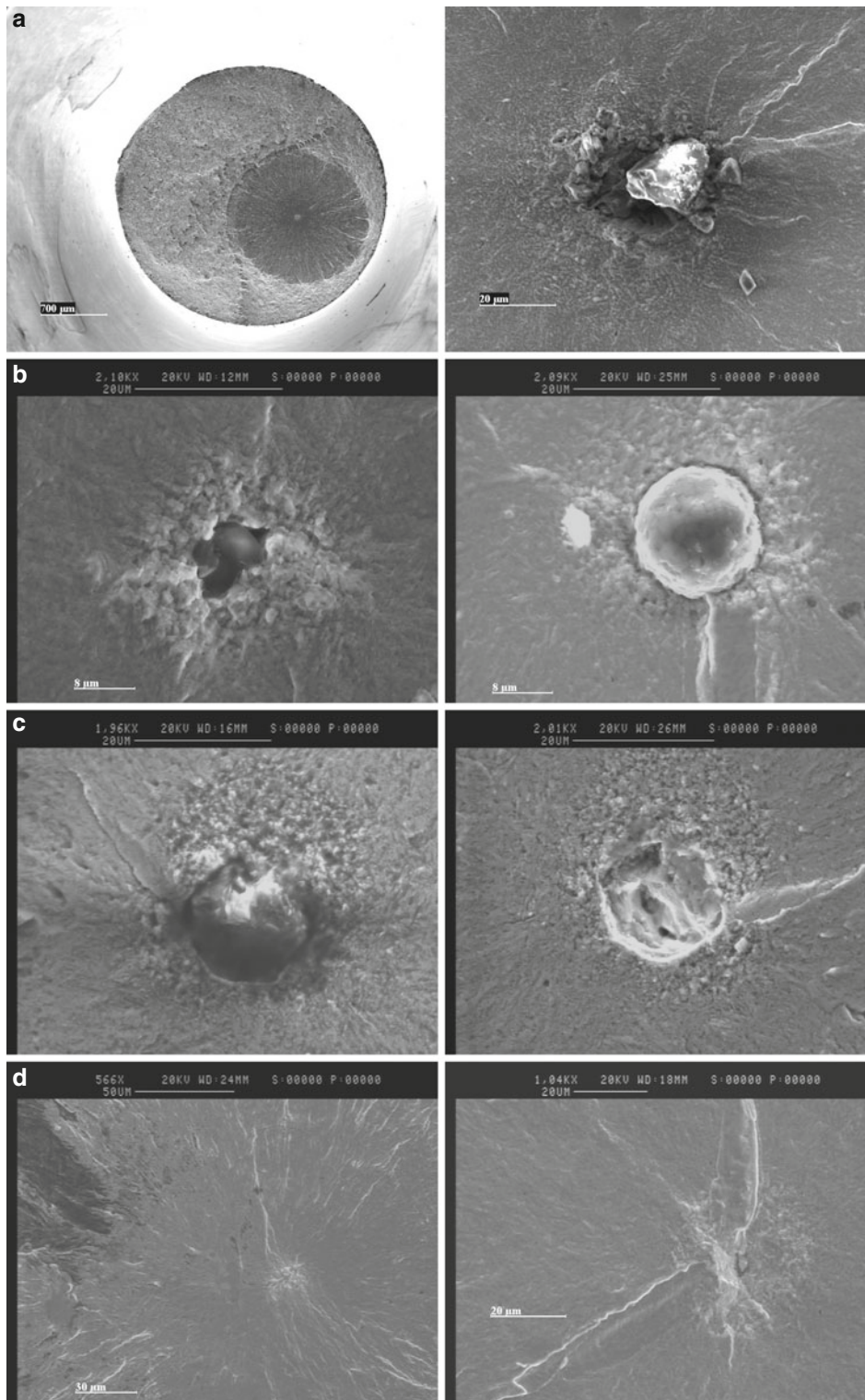
$$N_i = \frac{AW_s}{(\Delta\tau - 2\tau_f)^2} \quad (1)$$

where W_s is the specific fracture energy, $\Delta\tau$ is the range of local shear stress, and τ_f is the friction stress that needs to be overcome for the dislocations to move; A depends on the type of cracks initiated. The function A can be written as:

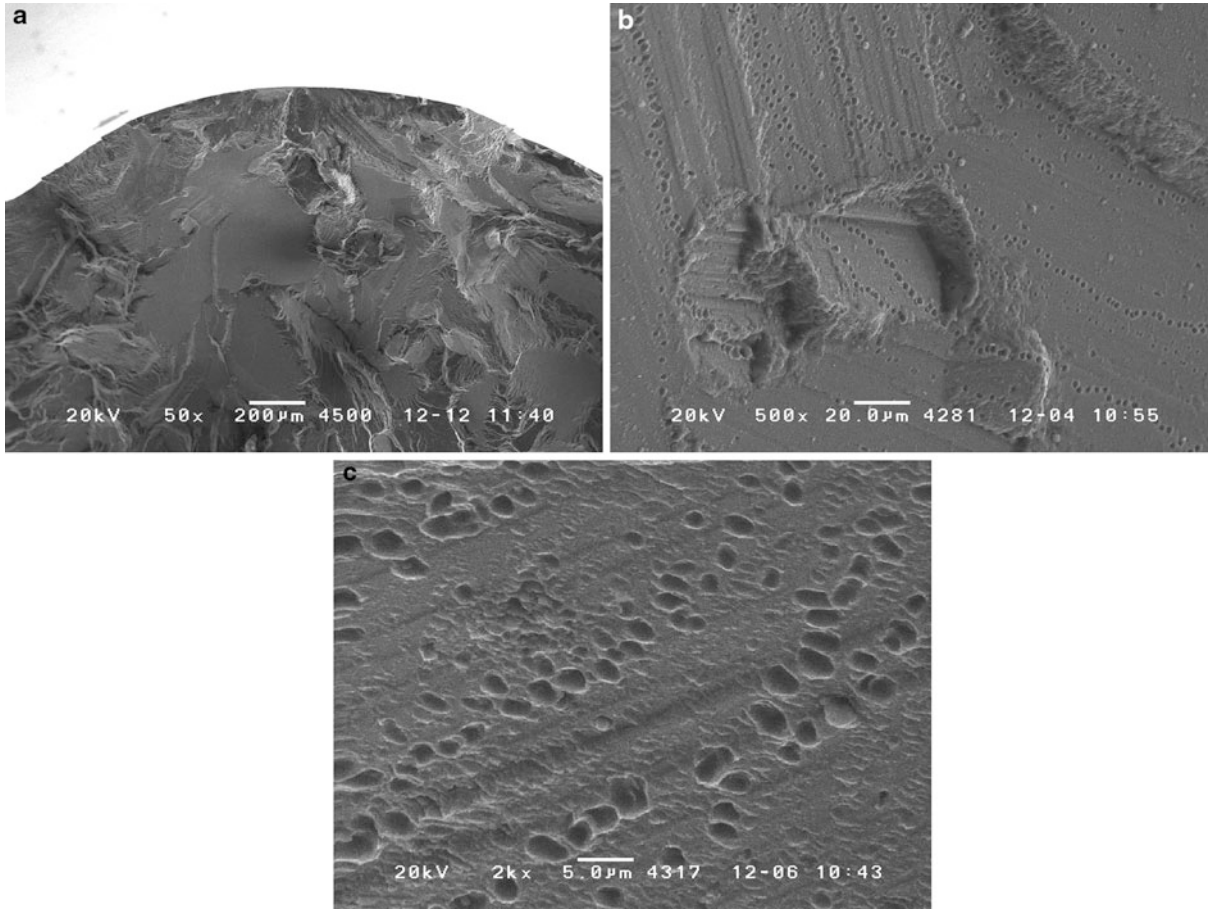
$$A = \begin{cases} \frac{4G}{\pi(1-\nu)l} & \text{Crack initiation along slip bands} \\ \frac{2G}{l} & \text{Crack initiation along grain boundary} \\ \frac{4G(G+G_i)h^2}{G_i(h+l)^2 a_i} & \text{Crack initiation along the interface of inclusion} \end{cases}$$

where G is the bulk shear modulus, G_i is the shear modulus of inclusion, l is the semi-length of slip band, h is the semi-minor length of the elliptical slip band area, and ν is the Poisson's ratio. Further, the following assumptions are made:

1. The crack initiation size is considered to be the initial inclusion size a_0 .
2. The semi-length of the slip band is assumed to be equal to the semi-minor length of the elliptical slip band area.



Super-Long Life Fatigue, Fig. 8 Some typical subsurface crack initiations



Super-Long Life Fatigue, Fig. 9 SEM micrographs showing (a) the fatigue fracture surface of 6061/T6 alloy and a drawing line distribution of fatigue voids (b) at low magnification (c) at high magnification (Roth and Roth 1987)

3. The shear modulus of inclusion is assumed to be equal to the bulk shear modulus.
4. The frictional stress τ_f is regarded as being half the fatigue limit, that is, no crack initiates at stress ranges lower than $2\tau_f$.

A physical interpretation of the von Mises yield criterion is that yielding occurs when the resolved shear stress on the octahedral plane exceeds the octahedral shear strength τ_o , given by

$$\tau_o = \frac{\sqrt{2}}{3} \sigma_o \quad (2)$$

where σ_o is the applied tensile stress.

The relationship between the shear stress amplitude and the applied stress amplitude is therefore defined as

$$\Delta\tau = \frac{\sqrt{2}}{3} \Delta\sigma \quad (3)$$

The friction stress is given by

$$\tau_f = \frac{1}{2} \left(\frac{\sqrt{2}}{3} \sigma_D^R \right) \quad (4)$$

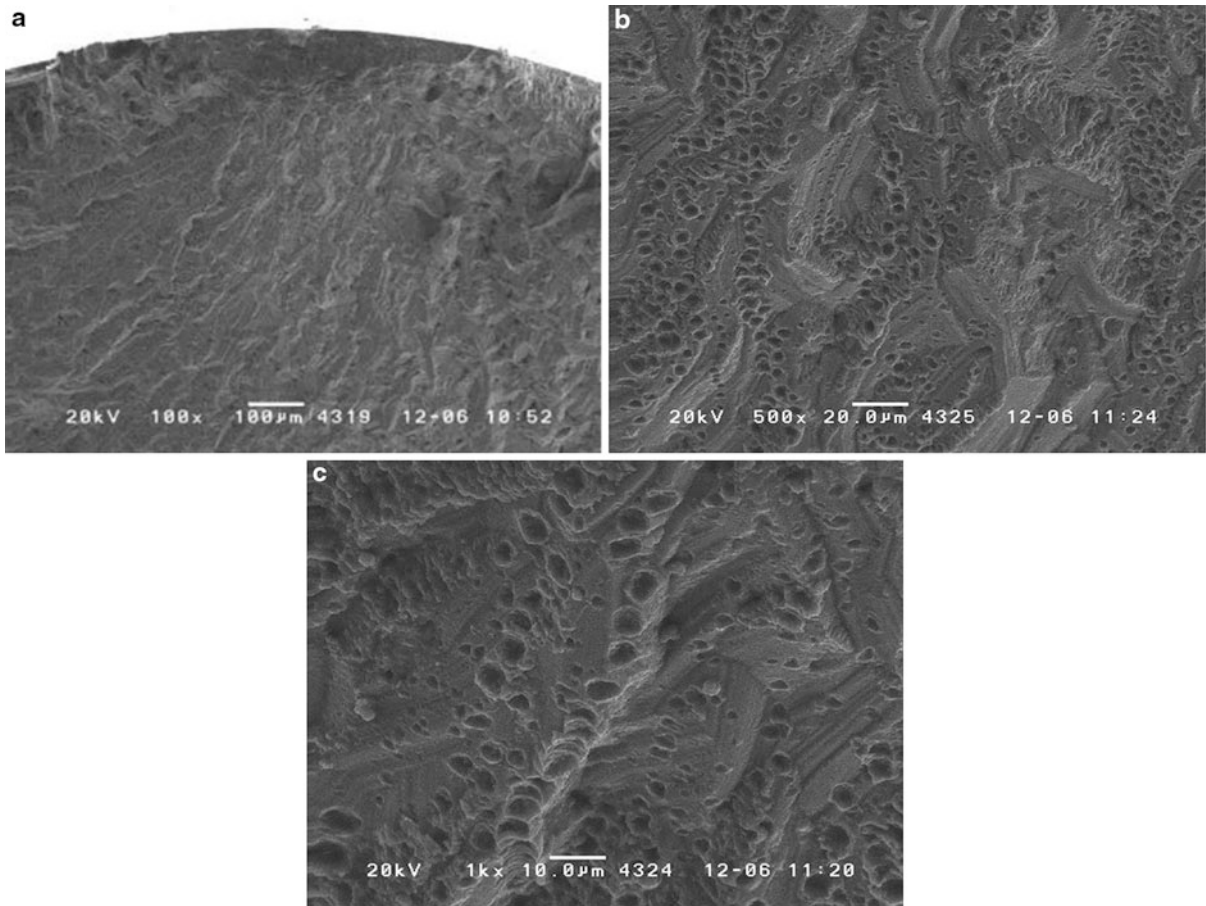
where σ_D^R is the fatigue limit of the material at a stress ratio of R.

The FCI life is thus calculated as

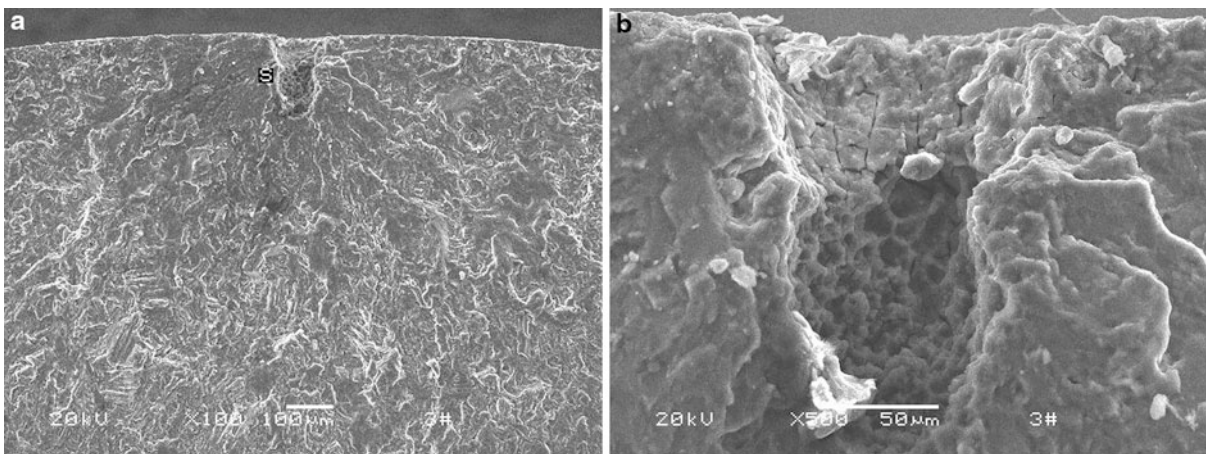
$$N_i = \frac{9GW_s}{(\Delta\sigma - \Delta\sigma_D^R)^2 a_0} \quad (5)$$

The specific fracture energy W_s can be obtained using Griffith crack theory (Anderson 1991), where the release energy due to crack propagation is equal to the energy consumed by new surfaces creation:

$$W_s = \frac{G_c}{2} = \frac{\Delta K_{th}^2}{2E} \quad (6)$$



Super-Long Life Fatigue, Fig. 10 SEM micrographs showing (a) the fatigue fracture surface of 7075/T6 alloy and fatigue voids (b) at low magnification (c) at high magnification (Wang et al. 2006)



Super-Long Life Fatigue, Fig. 11 SEM fractograph showing fracture origin of AE42 specimen tested to failure at 56.8 MPa, 2.1×10^8 cycles. (b) Higher magnification of (a)

where G_c is the critical elastic energy release rate and ΔK_{th} is the fatigue crack growth threshold.

Prediction of Fatigue Crack Growth Life

The number of cycles required for fatigue crack growth (FCG) from the point of initiation at the inclusion (of critical size) up to failure is modeled using Paris' law, as follows:

$$\frac{da}{dN} = C(\Delta K)^n \quad (7)$$

where a is the crack length, N is the number of cycles, and C and n are material properties.

The FCG life, considering both long and short crack (sc) behavior, is then determined by

$$N_p = \frac{a_o^{(1-\frac{n}{2})} - a_{sc}^{(1-\frac{n}{2})}}{C\Delta\sigma^n\beta_1^n\pi^{\frac{n}{2}}(\frac{n}{2}-1)} + \frac{a_{sc}^{(1-\frac{n}{2})} - a_f^{(1-\frac{n}{2})}}{C\Delta\sigma^n\beta_2^n\pi^{\frac{n}{2}}(\frac{n}{2}-1)} \quad (8)$$

Generally, $a_o \ll a_{sc}$ and $n > 2$, and the number of crack growth cycles from to is much greater than the cycles needed to grow from to final failure, particularly in high cycle fatigue. Then, N_p can be written as

$$N_p = \frac{a_o^{(1-\frac{n}{2})}}{C\Delta\sigma^n\beta_1^n\pi^{\frac{n}{2}}(\frac{n}{2}-1)} \quad (9)$$

Where β is the geometry constant equal to $0.5\sqrt{\pi}$.

Prediction of Total Fatigue Life

The total fatigue life N_f is the sum of the crack initiation life, N_i , and the crack growth life, N_p . Thus,

$$N_f = \frac{9GW_s}{(\Delta\sigma - \Delta\sigma_D^R)^2 a_o} + \frac{a_o^{(1-\frac{n}{2})}}{C\Delta\sigma^n\beta_1^n\pi^{\frac{n}{2}}(\frac{n}{2}-1)} \quad (10)$$

The above model was verified for Cr-Si steels (Wang et al. 2002). A fair agreement between the experimental values and the predicted values is seen in a few cases. It is further predicted that more than 99% of the fatigue life of Cr-Si steels in the super-long life regime could be attributed to the process of crack initiation. This seems to be the general observation among most researchers, as it has been widely accepted that in high cycle fatigue, in general, most of the fatigue life is consumed by the process of crack initiation.

Acknowledgments

Financial support from Chinese National Science Foundation through contract 10925211 is appreciated.

Cross-References

- Fatigue
- Fatigue Limit

- Fatigue Strength-Load Cycle Relationships for Ferrous Materials
- Growth Characteristics of Small Fatigue Cracks

References

- T.L. Anderson, *Fracture mechanics, fundamental and application* (CRC Press, Boca Raton, 1991)
- C. Bathias, C.P. Paris, *Gigacycle fatigue in mechanical practice* (Marcel Dekker, New York, 2004)
- D.L. Davidson, Fatigue crack growth at high R-ratio in Ti-6Al-4 V at 1.5 kHz: the effect of periodic removal of mean stress, in *Fatigue Behavior of Titanium Alloys*, ed. by R. Boyer, D. Eylon, J.P. Gallagher, G. Lütjering (TMS, Warrendale, 1999)
- R. Ebara, Y. Yamada, Ultrasonic corrosion fatigue testing of 13 Cr stainless steel and Ti-6Al-4 alloys, in *Ultrasonic Technology*, ed. by K. Toda (MYU Research, Tokyo, 1987), pp. 329–342
- S. Fukumoto, A. Mitchell, The manufacture of alloys with zero oxide inclusion content, in *Proceeding of the 1991 Vacuum Metallurgy Conference on the Melting and Processing of Specialty Materials 1 & SS*, Pittsburgh, 1991, pp. 3–7
- M. Kikukawa, K. Ohji, K. Ogura, *J. Basic Eng.* (Trans. ASME, D) **87**, 857 (1965)
- J.M. Morgan, W.W. Milligan, A 1 kHz servohydraulic fatigue testing system, in *High Cycle Fatigue of Structural Materials*, ed. by W.O. Soboyejo, T.S. Srivatsan (TMS-AIME, Warrendale, 1997), pp. 305–312
- Y. Murakami, T. Nomoto, T. Ueda, Y. Murakami, On the mechanism of fatigue failure in superlong life regime ($N > 10^7$ cycles). *Fatigue Fract. Eng. Mater. Struct.* **23**(11), 893–910 (2000)
- S. Nishijima, K. Kanazawa, Stepwise S-N curve and fish-eye failure in gigacycle fatigue. *Fatigue Fract. Eng. Mater. Struct.* **22**, 601 (1999)
- R. Prasannavenkatesan, J.X. Zhang, D.L. McDowell, G.B. Olson, H.J. Jou, 3D modeling of subsurface fatigue crack nucleation potency of primary inclusions in heat treated and shot peened martensitic gear steels. *Int. J. Fatigue* **31**(1), 1176–1189 (2009)
- L.D. Roth, L.D. Roth, Ultrasonic fatigue testing, in *Metals Handbook [M]*, vol. 8, Ninthth edn. (ASM, Metals Park, 1987), pp. 240–257
- T. Sakai, M. Takeda, K. Shiozawa, Y. Ohi, M. Nakajima, T. Nakamura, N. Oguma, Experimental evidence of duplex S-N characteristics in wide life region for high strength steels, in *7th International Fatigue Congress (Fatigue '99)*, Beijing, 1999, pp. 573–578
- K. Shiozawa, T. Hasegawa, Y. Kashiwagi, L. Lu, Very high cycle fatigue properties of bearing steel under axial loading condition. *Int. J. Fatigue* **31**(5), 880–888 (2009)
- K. Tanaka, T. Mura, A dislocation model for fatigue crack initiation. *Trans. ASME J. Appl. Mech.* **48**, 97 (1981)
- O. Umezawa, K. Ishikawa, Phenomenological aspects of fatigue life and fatigue crack initiation in high strength alloys at cryogenic temperature. *Mater. Sci. Eng. A* **176**(2–1), 397–403 (1994)
- Q.Y. Wang, J.Y. Berard, A. Dubarre, G. Baudry, S. Rathery, C. Bathias, Gigacycle fatigue of ferrous alloys. *Fatigue Fract. Eng. Mater. Struct.* **22**(8), 667–672 (1999)
- Q.Y. Wang, C. Bathias, N. Kawagoishi, Q. Chen, Effect of inclusion on subsurface crack initiation and gigacycle fatigue strength. *Int. J. Fatigue* **24**(12), 1269–1274 (2002)
- Q.Y. Wang, N. Kawagoishi, Q. Chen, Effect of pitting corrosion on the very high cycle life fatigue behavior. *Scr. Mater.* **49**(7), 711–716 (2003)

Q.Y. Wang, N. Kawagoishi, Q. Chen, Fatigue and fracture behaviour of structural Al-alloys up to very long life regimes. *Int. J. Fatigue* **28**(11), 1572–1576 (2006)

H.Q. Xue, H. Tao, F. Montebault, Q.Y. Wang, C. Bathias, Development of a three-point bending fatigue testing methodology at 20 kHz frequency. *Int. J. Fatigue* **29**(9–11), 2089–2093 (2007)

Super-Long Life Fatigue (SLLF)

► Super-Long Life Fatigue

Superlow Friction

ALI ERDEMIR

Energy Systems Division, Argonne National Laboratory
Tribology Section, Argonne, IL, USA

Synonyms

Near frictionless; Superlubricity

Definition

Super low friction is defined as a sliding regime in which friction or resistance to sliding is extremely low. Friction coefficients in this regime are typically below 0.01. Such levels of friction coefficients are typical of hydrodynamically separated or magnetically levitated sliding surfaces where little or no solid-to-solid contact takes place. However, attaining friction coefficients of less than 0.01 between sliding surfaces that are in direct metal-to-metal contact is difficult due to the complex physical, chemical, and mechanical interactions that take place among such contacting surfaces. Nevertheless, some

special or synthetic materials can provide such levels of friction under certain sliding conditions (Erdemir and Martin 2007).

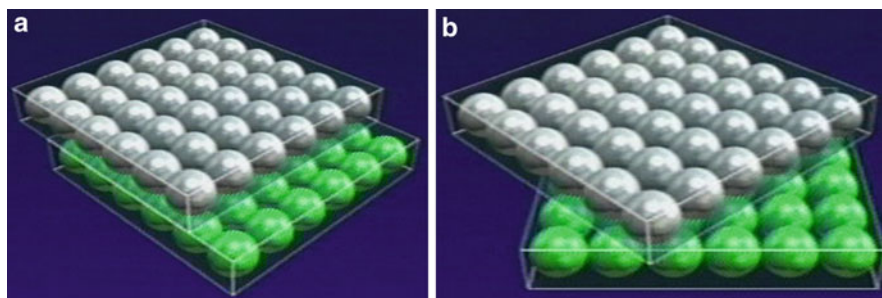
Scientific Fundamentals

Mechanisms

Theoretically, Sokoloff (1990) and Hirano and Shinjo (1990) predicted that super low friction or superlubricity should exist between atomically smooth surfaces of certain crystalline solids, provided they are brought into contact in an incommensurate or highly misaligned/misfit fashion. Having met such conditions, several researchers have observed super low friction between sliding incommensurate surfaces of graphite, MoS₂, and a few other solids (Martin et al. 1993; Hirano et al. 1997; Dienwiebel et al. 2004). Specifically, by rotating the surface atoms from a fully commensurate (Fig. 1a) to a completely incommensurate (Fig. 1b) state of contact, these researchers were able to show that a clear transition from pure stick-slip (indicative of high friction) to smooth sliding (indicative of no friction) occurred in their friction experiments. Figure 2 shows this transition for graphite.

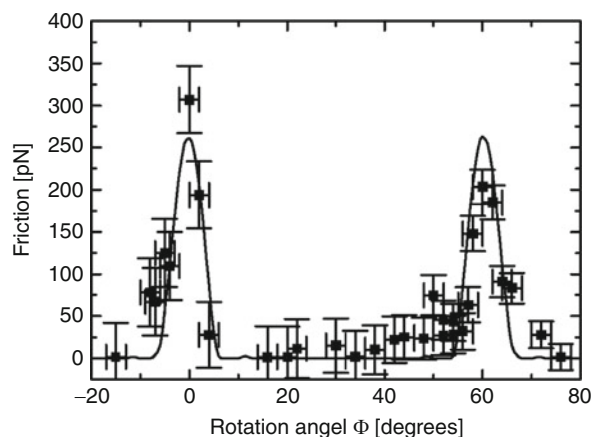
In addition to the crystalline solids mentioned, a new breed of diamond-like carbon (DLC) films was also shown to provide super low friction, even under severe contact conditions (Erdemir et al. 2000). These carbon films are synthesized in a highly hydrogenated gas discharge plasma using chemical and/or physical vapor deposition systems (Erdemir 2001). When tested in an inert or vacuum environment, these films provide friction coefficients of 0.001–0.005, as shown in Fig. 3.

When such DLC films are synthesized in a highly hydrogenated gas discharge plasma (in which the hydrogen-to-carbon atom ratio could be as high as 10), the density of hydrogen within the growing films as well as on their sliding surfaces becomes very high (≈40 at.%).



Superlow Friction, Fig. 1 Atomic-scale illustration of (a) commensurate (high friction) and (b) incommensurate (super low friction) contacts (Courtesy of Prof. Motohisa Hirano, Gifu University-Japan)

From a tribological point of view, the existence of large amounts of hydrogen on sliding surfaces of DLC can effectively eliminate the possibility of any unoccupied σ - or covalent bonds remaining and potentially participating in

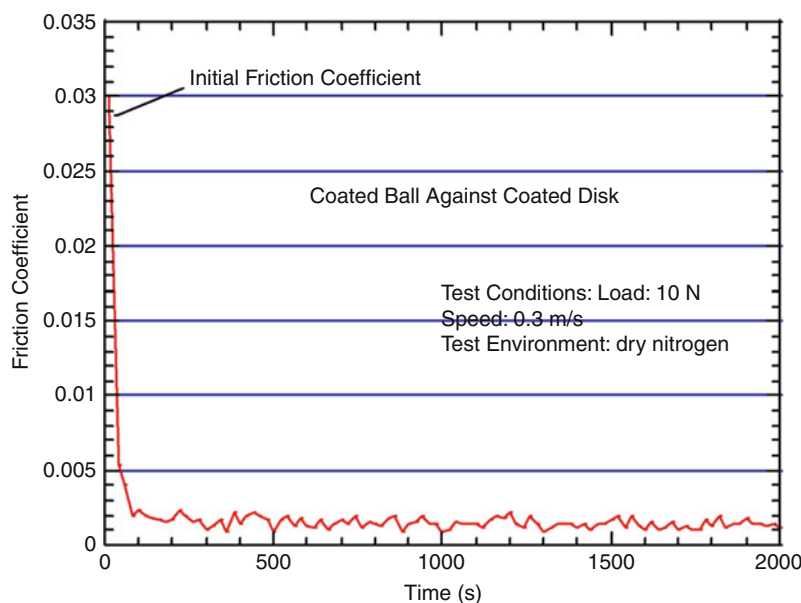


Superlow Friction, Fig. 2 Average friction force between a tungsten tip and a graphite substrate, plotted against rotation angle of the graphite sample. Two narrow peaks of high friction are observed at 0° and 61° , respectively, and these peaks represent a complete state of commensurability. Between these peaks are the regions with ultra-low friction due to incommensurate state of atomic-scale contact (Dienwiebel et al. 2004)

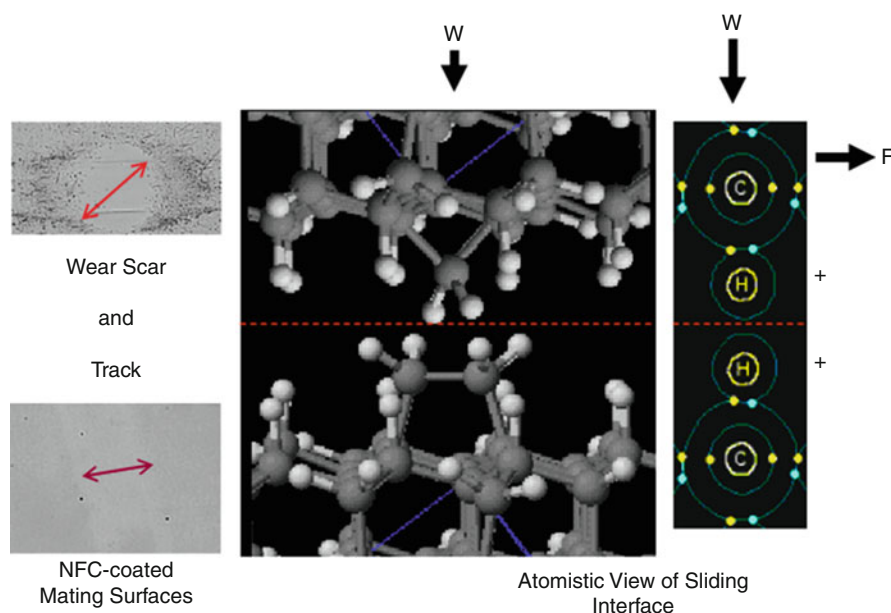
adhesive interactions during sliding (Erdemir 2001). Excess hydrogen within the film may always diffuse to the sliding surface and, hence, replenish or replace hydrogen that may have been removed due to frictional heating and/or mechanical wearing during sliding contact. As illustrated in Fig. 4, the free electrons of those hydrogen atoms on DLC surfaces are paired with the dangling σ -bonds of surface carbon atoms. As a result, the electrical charge density of hydrogen atoms is permanently shifted to the other side of the hydrogen nucleus and, hence, away from the surface. Such a situation will result in positively charged hydrogen protons being closer to the surface than the paired electron, which is tied to the dangling σ -bond of the surface carbon atoms. The creation of such a dipole configuration at the sliding interface may indeed give rise to repulsive rather than attractive forces among the hydrogen-terminated sliding surfaces of the DLC films. In support of the proposed mechanism described above, using computer simulations, Dag and Ciraci (2004), Qi et al. (2006), and Schall et al. (2010) have recently demonstrated the existence of such repulsive forces between H-terminated diamond (001) and DLC surfaces.

Super Low Friction by Other Solids

Due to their lamellar structures, graphite and a few inorganic solids (MoS_2 , H_3BO_3 , etc.) can provide very low friction and wear coefficients, depending on test



Superlow Friction, Fig. 3 Super low friction behavior of a highly hydrogenated DLC film in dry nitrogen



Superlow Friction, Fig. 4 Atomic-scale depiction of highly hydrogenated DLC surfaces brought into contact

environments and conditions. For example, when tested in inert gases or vacuum, MoS_2 was shown to provide friction coefficients of less than 0.01. Martin et al. (1993) achieved superlubricity in ultra-pure and thin MoS_2 films under ultrahigh vacuum conditions. Hirano et al. reported by factors of up to four reductions in the amount of frictional force when two mica sheets were increasingly rotated out of registry and finally brought into an incommensurate state of sliding (Hirano et al. 1991). Hirano's group has also demonstrated super-low friction for tungsten tip against silicon wafer under ultra-high vacuum conditions (Hirano et al. 1997). Likewise, Dienwiebel et al. (2004) and Mate et al. (1987) observed near-zero friction between sliding surfaces consisting of tungsten tips and graphite sheets at nano-scale contacts. In Dienwiebel's work, achieving superlubricity required complete incommensurability as a precondition (as is clear from Fig. 2) along with testing in ultra-clean test environments. In a related study involving graphite (Miura et al. 2005), nearly frictionless sliding regimes were also achieved on C_{60} -intercalated graphite films in which alternating monolayers of C_{60} were introduced between the graphene sheets. It was assumed that C_{60} was not only increasing the inter-planar spacing but also acting as molecular-scale ball bearings. Super low friction appears to be not limited to the lamellar solids mentioned above. Studies by Masuda and Honda (2003) have also

shown that hydrogen-terminated $\text{Si}(111)$ surfaces can attain super low friction (0.003) when rubbed against a diamond surface in ultra-high vacuum.

Systematic lubrication studies by Kano et al. (2005) have resulted in extremely low friction coefficients on hydrogen-free DLC films under lubricated sliding conditions. Specifically, they blended a poly-alpha-olefin base oil with glycerol mono-oleate (GMO) and used it as a boundary lubricant for sliding surfaces of hydrogen-free DLC films. Surface analytical studies confirmed the presence of a layer of OH on rubbing surfaces. Based on these findings, Kano et al. proposed that alcohol function groups of GMO and the surface carbon atoms of tetrahedral amorphous carbon (ta-C) were mechanically and tribochemically activated to result in a strongly bonded OH layer on ta-C surfaces. Just like hydrogen termination, OH termination of the dangling σ -bonds appears to have resulted in such ultralow friction. In another related study, Kato and his coworkers (2003) demonstrated the feasibility of achieving super low friction on a specially prepared carbon nitride (CN_x) film. They obtained such low friction coefficients by blowing nitrogen gas onto the sliding surfaces of Si_3N_4 balls rubbing against the CN_x coatings.

Not so surprisingly, in most sliding systems where superlubricity was observed, the amount of wear was also extremely small or difficult to quantify. It appears that within the super lubricated sliding regime,

the surfaces are able to slide against one another without creating much wear.

Key Applications

Friction is a fascinating, challenging, and important phenomenon. In daily life, friction is extremely important for safety and mobility. It greatly impacts energy efficiency and durability in all kinds of mechanical systems. Articular joints work well and last long thanks to the super low friction nature of smooth cartilages lubricated with a synovial fluid – allowing nearly frictionless movement. In many of the other moving mechanical assemblies, super low friction is desired to reduce energy losses due to friction. In car engines, 10–15% of the fuel's energy is lost to overcome friction; burnt fuel turns into CO₂ and goes into the atmosphere, exacerbating global warming. MoS₂ is used extensively to combat friction and wear in numerous space applications. Nano-sheets, tubes, and fullerene-like nanomaterials made out of MoS₂, WS₂, H₃BO₃, hexagonal boron nitride, graphite, and carbon are being considered as nano-colloidal additives for a variety of lubricants in an attempt to further enhance their anti-friction and -wear properties (Martin and Ohmae 2008).

In recent years, DLC films have become one of the most valuable engineering materials for friction and wear control in a number of industrial applications, including microelectronics, manufacturing, transportation, and biomedical fields. In magnetic storage media, DLC films have been used for a long time. Over the years, the uses of DLC have increased in many other sectors thanks to the introduction of industrial-scale, more robust coating systems that can produce high-quality DLC films on all kinds of substrates. Since the late 1990s, DLC films have been used extensively in razor blades and fuel injector systems of diesel engines. At present, high-quality DLC films are readily available from many commercial sources. Some of these DLC coatings are alloyed with other elements and others contain unique crystalline nanostructures and/or nano-phases that make them more durable and multi-functional to meet the increasingly more stringent application conditions of advanced mechanical devices. DLC films are now routinely produced on a variety of mechanical systems ranging from razor blades to micro-electromechanical systems, from numerous engine parts to articulated hip and knee joints, from bearings to machine tools and dies. Overall, the development and increased uses of super low-friction materials, coatings, and lubricants have very important implications for not only saving energy, but also protecting life from harmful emissions.

Acknowledgment

This work was supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Freedom Car and Vehicle Technologies Program, under Contract No. DE-AC02-06CH11357.

Cross-References

- [Basic Concepts in Adhesion Science](#)
- [Diamond-like Carbon Coatings](#)
- [Friction \(Concepts\)](#)

References

- S. Dag, S. Ciraci, *Phys. Rev. B* **70**, 241401 (2004)
- M. Dienwiebel, G.S. Verhoeven, N. Pradeep, J.W.M. Frenken, J.A. Heimberg, H.W. Zandbergen, *Phys. Rev. B* **92**, 126101 (2004)
- A. Erdemir, *Surf. Coat. Technol.* **146**, 292 (2001)
- A. Erdemir, J.-M. Martin (eds.), *Superlubricity* (Elsevier, Amsterdam, 2007)
- A. Erdemir, O.L. Eryilmaz, G. Fenske, *J. Vac. Sci. Technol.* **A18**, 1987 (2000)
- M. Hirano, K. Shinjo, Atomistic locking and friction. *Phys. Rev. B* **41**(17), 11837–11851 (1990)
- M. Hirano, K. Shinjo, R. Kaneko, Y. Murata, *Phys. Rev. Lett.* **67**, 2642 (1991)
- M. Hirano, K. Shinjo, R. Kaneko, Y. Murata, *Phys. Rev. Lett.* **78**, 1448 (1997)
- M. Kano, Y. Yasuda, Y. Okamoto, Y. Mabuchi, T. Hamada, T. Ueno, J. Ye, S. Konishi, S. Takeshima, J.M. Martin, M.I.D. Bouchet, T. Le Mogne, *Trib. Lett.* **18**, 245 (2005)
- K. Kato, N. Umehara, K. Adachi, *Wear* **254**, 1062 (2003)
- J.M. Martin, N. Ohmae, *Nanolubricants* (Wiley, West Sussex, 2008)
- J.M. Martin, C. Donnet, T.H. Le Mogne, T.H. Epicier, *Phys. Rev. B* **48**, 10583 (1993)
- H. Masuda, F. Honda, *IEEE Trans. Magn.* **39**, 903 (2003)
- C.M. Mate, G.M. McClelland, R. Erlandsson, S. Chiang, Atomic-scale friction of a tungsten tip on a graphite surface. *Phys. Rev. Lett.* **59**(220), 1942–1945 (1987)
- K. Miura, D. Tsuda, N. Sasaki, *J. Surf. Sci. Nanotechnol.* **3**, 21 (2005)
- Y. Qi, E. Konca, A.T. Alpas, *Surf. Sci.* **600**, 2955 (2006)
- J.D. Schall, G. Gao, J.A. Harrison, *Phys. Chem. C* **114**, 5321 (2010)
- J.B. Sokoloff, *Phys. Rev. B* **42**, 760 (1990)

Superlubricity

- [Superlow Friction](#)

Surface Adsorption in CVD

- [Chemical Vapor Deposition Processes for Boundary Lubricants](#)

Surface Analysis Using Contact Mode AFM

MIGUEL ANGEL SÁNCHEZ QUINTANILLA

Dpto. Electrónica y Electromagnetismo, University of Seville, Seville, Spain

Synonyms

Contact Mode Scanning Force Microscopy (SFM)

Definition

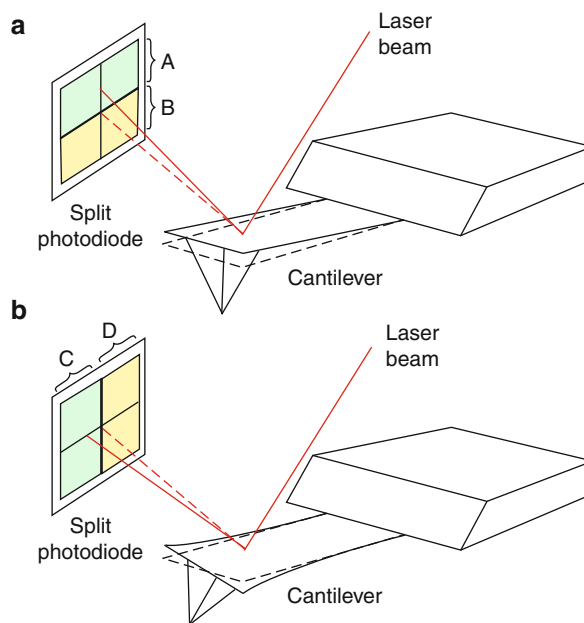
The atomic force microscope (AFM) is an apparatus used to obtain high-resolution three-dimensional images of surfaces by monitoring the interaction between a mechanical probe and the imaged surface. Different modes of operation are possible depending on the kind of interaction between probe and surface. In particular, in contact mode the probe is kept in contact with the surface.

Scientific Fundamentals

Contact mode AFM was invented by Gerd Binnig and Christoph Gerber in 1986 combining the principles of the scanning tunneling microscope and stylus profilometry. In an AFM the probe is a rectangular or V-shaped micrometer-sized lever usually referred to as a “cantilever.” A sharp tip is placed at the free end of the cantilever. When Binnig and his co-workers published their invention, they suggested that this new type of microscope would be able to measure forces between individual atoms, and accordingly they called it the atomic force microscope (AFM). However, in most situations the force between the pair of closest atoms in the tip and the substrate does not account for the total of the interaction force between the cantilever’s tip and the substrate and therefore the term scanning force microscope is also used to refer to this type of microscope.

Principle of Operation

In contact mode AFM, the cantilever’s tip is kept in contact with the surface during image acquisition. The cantilever and the surface are pressed against each other and the repulsive force arising from the interaction of the atoms of the tip’s apex with those of the surface cause a deflection of the cantilever. In their prototype, Binnig and co-workers used a scanning tunneling microscope to monitor the deflection of the cantilever. However, most of contemporary AFMs rely on the use of an optical lever, in which a laser beam is focused on the free end of the cantilever. The beam reflected from the cantilever is directed to a split



Surface Analysis Using Contact Mode AFM, Fig. 1 Principle of operation of the optical lever to detect: (a) cantilever deflection; (b) cantilever torsion

photodiode. In a simple version of the optical lever the photodiode is divided in two sections, A and B, and the path of the reflected beam is arranged in such a way that a deflection of the cantilever causes the incidence spot of the laser beam on the photodiode to move perpendicular to the line separating the sections A and B, as shown in Fig. 1. If the center of the incidence spot is placed on this line before the cantilever is pressed against the substrate, the difference in voltage output A-B gives a measure of the cantilever deflection. More elaborate versions of the optical lever use a photodiode divided in four sections by an additional line parallel to the direction of the displacement of the laser spot caused by a deflection of the cantilever. The voltage difference C-D between the sections of the photodiode on either side of this second line is related to the torsion of the cantilever. In short, split photodiodes are only capable of determining the deflection of the cantilever, while quadruple photodiodes can measure both the deflection and the torsion of the cantilever.

The sample to be imaged is mounted on a scanner consisting of a tube with vertically segmented walls, each of the segments being a piezo-electric. By applying voltages to the appropriate segments, the scanner can move the sample in both horizontal directions (XY) and in the vertical direction (Z). The alternative design, in which the

cantilever, rather than the sample, is mounted on a piezoelectric scanner, is also used by some AFM models, but since the operation of the AFM is similar in both designs, it will not be considered here.

During image acquisition, the sample is rastered against the cantilever tip in parallel lines, much like the raster of the image in a cathode ray tube television. By convention, the direction of the lines is called X-axis or “fast scan axis” while the direction perpendicular to it is the Y-axis or “slow scan axis.”

Resolution

A common feature of all AFM images is that the vertical resolution is greater than the horizontal resolution. The vertical resolution of an image is limited by the scanner properties, the noise level in the A-B signal (usually due to vibrations introduced by the environment’s influence), and the cantilever normal stiffness. Under optimal conditions, contact mode AFM can achieve a vertical resolution below 0.1 nm. In order to achieve good vertical resolution it needs to have a large deflection of the cantilever with a minimum force. This requires a cantilever that is as soft as possible. For this reason, the normal stiffness K_n of contact mode cantilevers are one to two orders of magnitude lower than the cantilevers used in other modes of operation. For example, typically K_n is in the range of 0.05–5 N/m for contact mode cantilevers compared with typical values of K_n in the range of 5–50 N/m for tapping mode cantilevers.

In contrast, given a scanner with an adequate horizontal range, the horizontal resolution is limited by the sharpness of the tip. Assuming the tip and the substrate interact as perfectly rigid bodies, the maximum attainable vertical δz_v and horizontal δz_h resolution are related by the expression:

$$\delta z_h = \sqrt{2R\delta z_v} \quad (1)$$

where R is the tip radius. Most nominal conventional tip’s radius range from 10 to 40 nm. Assuming $\delta z_v = 0.1$ nm, $R = 10$ nm, $\delta z_h = 1.41$ nm, meaning that very sharp tips are needed to obtain sub-nanometer horizontal resolution. In fact, this is just an overestimation of the maximum horizontal resolution, since under the combined effect of the load force and the adhesion both cantilever and substrate experience surface deformations that reduce the horizontal resolution. In practice atomic resolution is only achieved by contact mode AFM when very flat substrates are imaged in liquid, because in this situation the adhesive forces between tip and substrate are minimized.

The aspect ratio of the tip must also be taken into account when measuring the size of small and/or sharp features in AFM images due to the artifact known as tip broadening (Braga 2004): the width of the cantilever tip at the feature’s height adds to the measured length in the XY plane. As tips wear during use, it is important to characterize the correct shape of the tip apex if accurate length measurement is pursued. Commercially available substrates presenting ridge- and wedge-shaped edges are commonly used as tip characterizers. They are based on an artifact known as “self-imaging” or “reverse imaging”: when the tip scans a feature whose curvature radius is smaller than the tip’s radius, the result is an image of the tip rather than one of the feature on the substrate.

Key Applications

The main advantage of the AFM over other high-resolution imaging techniques (scanning tunneling microscopy and scanning electron microscopy) is its capability to obtain images of non-conductive surfaces without the need of conductive coatings and its ability to operate with samples immersed in a liquid, for example, some in vivo biological samples such as cell cultures. Additionally, as the AFM works by responding to very small forces, it is also used to measure the adhesion between particles and substrates and for friction and wear studies at the microscopic scale (Bhushan and Fuchs 2009).

Unlike scanning electron microscope (SEM) images, AFM images contain full 3D topographical information of the substrate. Therefore, the tools of quantitative analysis devised for stylus profilometry such as line and area roughness, bearing curve, or spectral analysis are also applicable to AFM images.

The main drawbacks of the AFM with respect to the SEM are that it is difficult to obtain quality images of samples with a rough topography, as, for example, micrometer-sized particles or substrates with steep gradients in height, and that image acquisition is usually much more time consuming than in SEM because the image refresh rate is much lower. Typical scanning rates in contact mode AFM are in the range of 1–5 lines per second, which amounts to a total of 1.7–8.5 min to obtain a 512-line image.

Because of this slow image refresh rate, considerable time may be spent searching for interesting features in the substrate (which can be minimized by setting less lines per image while searching) or distortions due to thermal drift may appear in the image. These reasons, combined with the desire to image fast-occurring processes in real time, have prompted research in the so-called high-speed contact mode AFM. This type of AFM can produce images

rates of frames per second, comparable to the image refresh rate of scanning electron microscopes. High speed AFMs differ by their construction from conventional AFMs since in conventional AFMs the scan rate is limited by the mechanical resonances of the piezo-electric scanner, the inertia of the cantilevers, and the processing speed of the feedback loop electronics. The existing designs of high-speed AFMs differ in the strategies used to overcome these limitations.

Scanning Modes

Topography

Two scanning modes to obtain topographical information are possible in contact mode: constant height or constant deflection. In constant height, the vertical extension of the piezo-electric scanner is kept constant as the sample is rastered under the cantilever, the topography of the sample causing a position-varying deflection of the cantilever. The image is formed by plotting the A-B signal against the horizontal displacements of the scanner. Due to the fact that the cantilever may break should it encounter a prominent feature on the surface, constant height images are not recommended if the surfaces are not very flat. In constant deflection, a feedback loop acting on the vertical elongation of the piezoelectric (Z-signal) and fed with the A-B signal adjusts the Z-signal to maintain the A-B signal constant at a value chosen by the user. This value is known as the setpoint and it is related to the normal force applied to the surface by the cantilever's tip during the scan. The image is formed by plotting the vertical extension of the piezo-electric scanner (Z-signal) versus the horizontal displacements of the scanner, as in Fig. 2. Another image can be obtained by plotting the difference between the A-B signal and the setpoint versus the horizontal displacements (error signal image). If the feedback loop is properly tuned, the A-B signal only differs from the setpoint briefly when the cantilever passes over a strong variation in the topography of the surface (see, for example, Fig. 3). Because of this, the error signal image reflects the gradients in the height of the sample and shows sharp features as cracks, borders, or spikes more neatly than the topographical image.

Constant deflection has the advantage over constant height that a constant force is applied to the sample during image acquisition. As has been mentioned, this force is related to the setpoint. In order to transform the setpoint from units of voltage to units of force it is required to know both the normal sensitivity of the optical lever S_n and the normal stiffness of the cantilever K_n . The normal sensitivity S_n is defined as the constant relating the change

in the voltage output ΔV_{A-B} of the photodiode to the vertical deflection of the cantilever Δz_c :

$$\Delta V_{A-B} = S_n \Delta z_c \quad (2)$$

The normal stiffness relates the vertical deflection of the cantilever with the normal force F_n acting on it:

$$F_n = K_n \Delta z_c \quad (3)$$

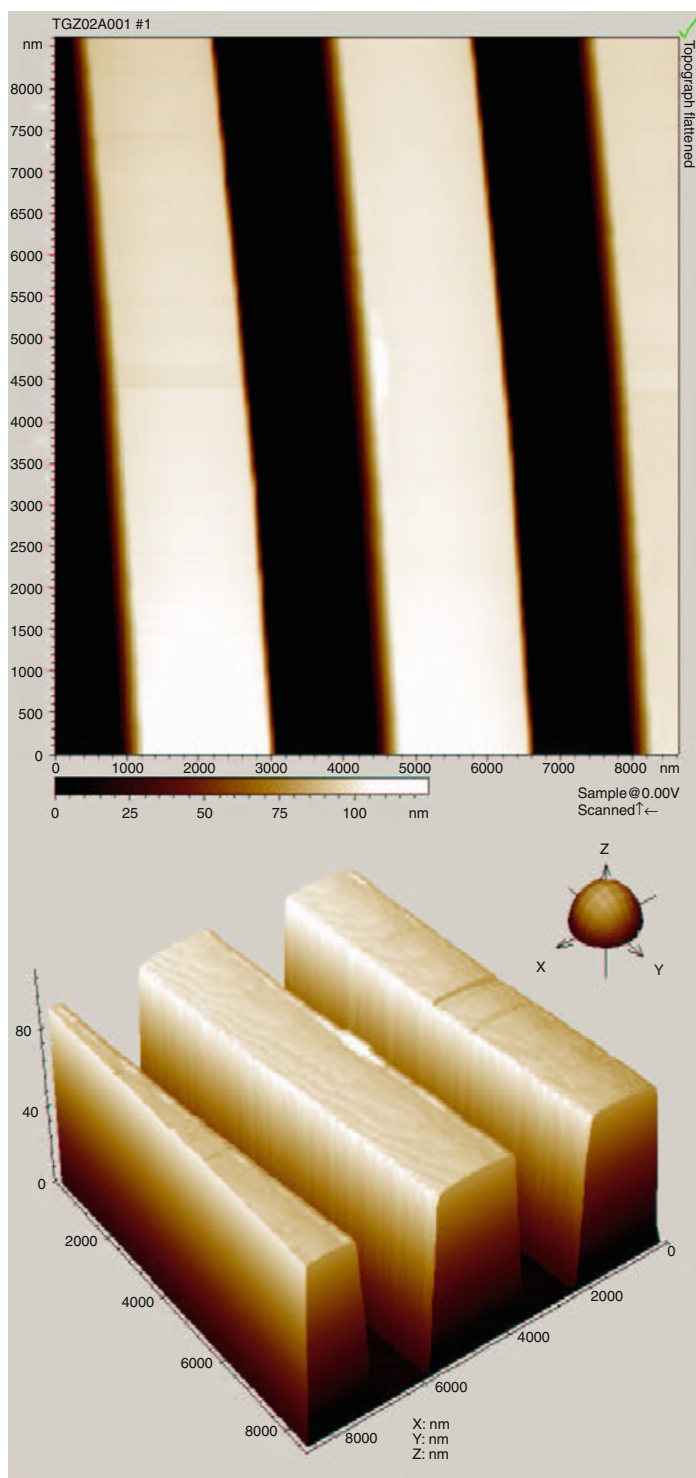
Thus, if an image was acquired with a setpoint of V_0 volts, the repulsive force between the tip and the surface during the image acquisition was of $F_n = K_n V_0 / S_n$ nN if K_n is given in nN/nm and S_n in V/nm.

Usually, to obtain an image with setpoint V_0 the AFM is operated in such a way that the voltage output V_{A-B} of the optical lever equals $-V_0$ when the cantilever is unengaged and requires a sample approach such as the voltage output V_{A-B} of the optical lever is equal to zero when the cantilever's tip is in contact with the surface (thus $\Delta V_{A-B} = V_0$). The reason for this is that the optimal linearity in the response of the optical lever is obtained when the center of the laser spot is on the line dividing the A and B sections of the photodiode.

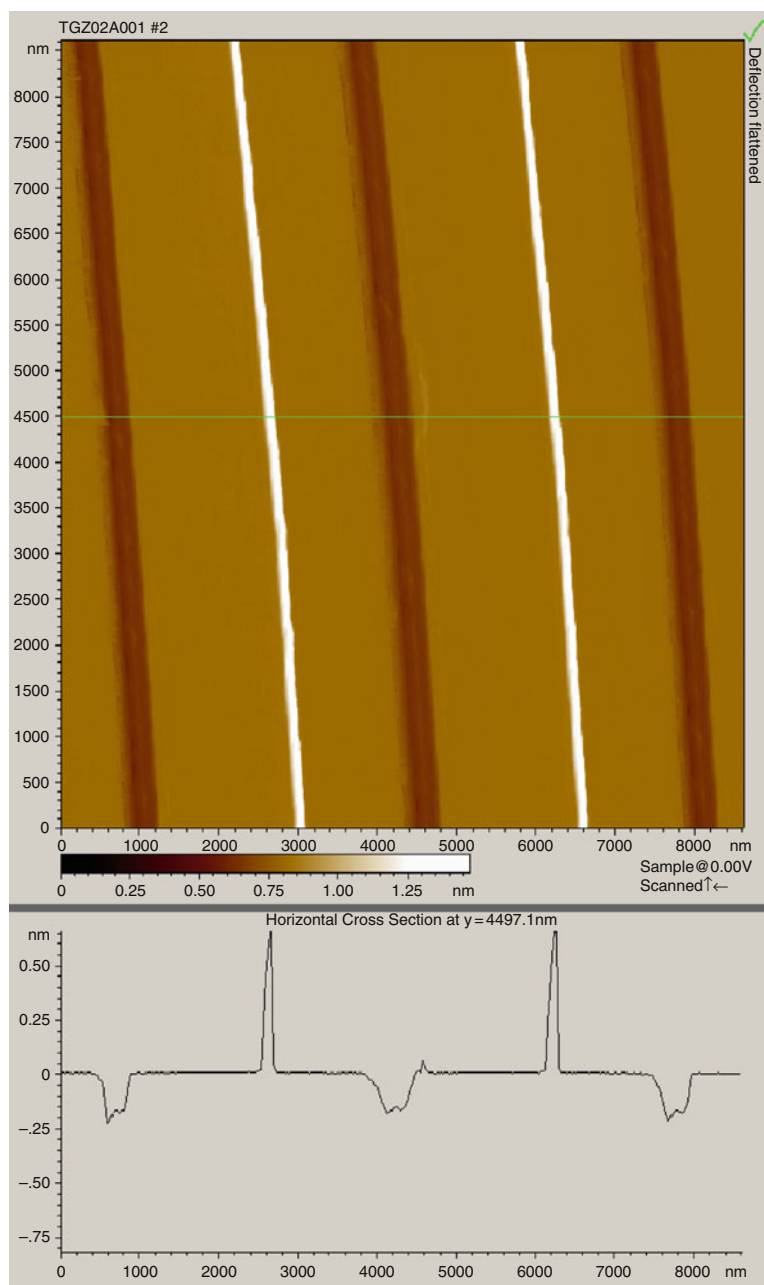
There are occasions in which varying the setpoint has an effect over the image: for example, when the scanned surface has holes or crevices, increasing the setpoint forces the cantilever's tip to dip deeper into those features, revealing their inside, which otherwise would have appeared as a featureless area in the image. However, increasing the setpoint increases the likelihood of damage both to the surface or the cantilever's tip during image acquisition.

Friction

As the cantilever has to slide over the substrate during image acquisition, a frictional force acts on the cantilever's tip. The cantilever's mode of deformation in response to the friction force depends on the orientation of the fast scan axis with respect to the long axis of the cantilever. Usually, this orientation is a user-selectable parameter in the image acquisition software of the microscope. When the fast scan axis is aligned with the cantilever's long axis, the friction force causes an additional bending of the cantilever, whereas a perpendicular alignment causes a torsion of the cantilever around its long axis. As has been said, the torsion of the cantilever causes a change in the C-D signal in a properly aligned quadruple photodiode detector. Lateral force images such as the one shown in Fig. 4 consist of a representation of the C-D signal versus the horizontal displacement of the scanner acquired simultaneously with a constant deflection topographical image. Lateral force images show



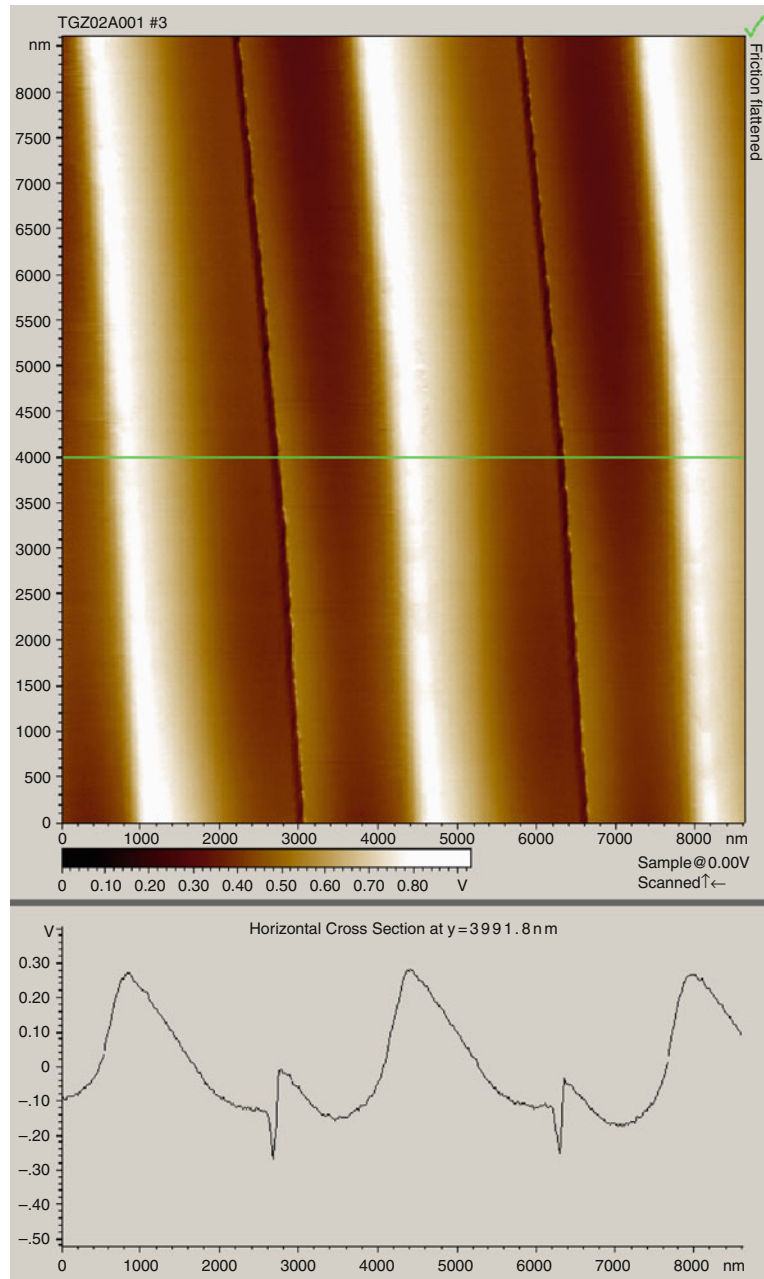
Surface Analysis Using Contact Mode AFM, Fig. 2 Topograph image of a calibration grating consisting on a 1D array of rectangular steps (*top*) and a 3D representation of the height data in the image (*bottom*). The image was acquired in constant deflection mode using the PicoScan 2500 AFM at CITIUS (University of Seville) with a PPP-NCH-20 cantilever (force constant ~ 45 N/m)



Surface Analysis Using Contact Mode AFM, Fig. 3 Deflection or error signal image acquired simultaneously with the topographical image shown in Fig. 2 (top) and a scan line at the position of the horizontal line (bottom). Note that the cantilever's deflection is the same in the pits and the heights of the steps and corresponds to the imaging setpoint (0 V, the vertical units in the image have been transformed into nm using the normal sensitivity – see under "Sensitivities" below) and that it only differs briefly when the cantilever is climbing up (upward deflection) or down a step (downward deflection). The cantilever moves in the scan line shown at the bottom from right to left

regions of different frictional characteristics on the substrate. However, obtaining quantitative information about the friction coefficient μ between tip and sample requires

knowing the torsional stiffness K_ϕ of the cantilever and the lateral sensitivity S_l of the optical lever. The torsional stiffness relates the torque acting on the cantilever along



Surface Analysis Using Contact Mode AFM, Fig. 4 Friction signal image acquired simultaneously with the topographical image shown in Fig. 2 (top) and a scan line at the position of the horizontal line (bottom). In this case, the friction signal arises from the torsion of the cantilever when its tip falls down a step wall (broad peaks) or collides with it (sharp valleys). The cantilever moves in the scan line shown at the bottom from right to left and the scan direction is perpendicular to the cantilever's long axis

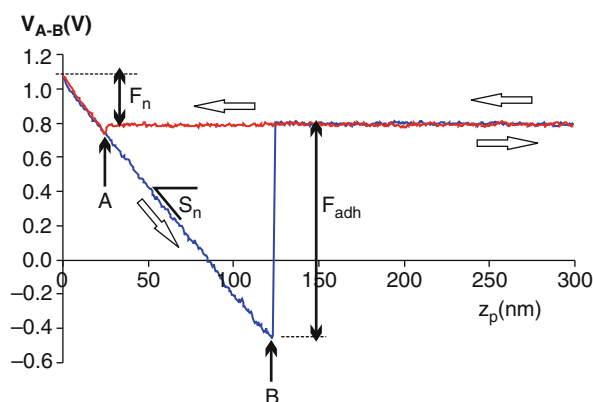
the cantilever's long axis with the angle of torsion $\Delta\phi$ and the lateral sensitivity S_l is the constant that relates the change in the voltage output ΔV_{C-D} of the photodiode with the change in the angle of torsion $\Delta\phi$: $\Delta V_{C-D} = S_l \Delta\phi$.

Force Spectroscopy: Load Curves

In most circumstances, there is a force of adhesion between the cantilever and substrate when they are in contact. There are a number of possible sources of the

adhesion between tip and sample: capillary forces due to absorbed liquid layers, electrostatic forces, or van der Waals forces are among the most frequently cited. To determine the nature of the dominant forces acting between the tip and a specific substrate requires knowledge of the working conditions and the state and properties of the imaged surface.

The value of the adhesion between tip and sample can be measured monitoring the deflection of the cantilever as a function of the position of the sample (see Fig. 5). Such curves are called “load curves.” When acquiring a load curve, the feedback loop is disabled and the sample is pressed against the cantilever. Usually, when the distance between the cantilever’s tip and the surface is smaller than some certain distance, the tip snaps into contact with the substrate due to attractive forces acting between the substrate and the cantilever (point A in Fig. 5). From this point onwards, the tip and the substrate move together. In most cases, when the substrate is retracted, the cantilever’s tip does not separate from the substrate until some downward deflection of the cantilever (point B in Fig. 5). The reason for this is that the elastic force trying to return the cantilever to its neutral position must overcome the adhesion between the cantilever and the substrate.



Surface Analysis Using Contact Mode AFM, Fig. 5 A typical load curve obtained with a colloidal probe showing the A-B signal V_{A-B} as a function of the vertical position z_p of the substrate. The origin of z_p is arbitrary, z_p increasing means that the substrate moves to separate itself from the particle attached to the cantilever. Points A and B mark the point of first contact of the particle with the substrate and the point of detachment, respectively. The double-headed arrows labeled F_n and F_{adh} mark the voltage differences corresponding to the load force (F_n) and the adhesion F_{adh} . The slope S_n gives the normal photodiode sensitivity. The void arrows indicate the direction of the movement of the substrate along the curve

The magnitude of the adhesion is proportional to the jump in the A-B signal when the cantilever snaps out of contact.

Some information on the nature of the adhesive forces acting between tip and substrate is provided by the shape of the load curve: a sharp pull-off indicates the presence of short-range forces such as van der Waals forces, a gradual recovery followed by a sharp pull-off may be explained by the rupture of a liquid neck between tip and substrate, a gradual recovery after pull-off indicates the action of long-range forces such as electrostatic force, while multiple jumps during pull-off indicate multiple separation events as in the breakage of polymer chains.

Load curves acquired using modified cantilevers have been frequently used to measure the adhesion forces between two dry surfaces as a procedure to obtain surface energy values. The most usual method of cantilever modification consists of attaching a particle on a tipless cantilever (such modified cantilevers are sometimes referred to as “colloidal probe” cantilevers). However, determining with some degree of accuracy the surface energy requires resolving the topography of the contacting surfaces because the adhesion not only depends on the surface energy but also on the real area of contact, which in turns depends on the roughness of the contacting surfaces.

Force Calibration

One of the things that makes the atomic force microscope attractive is its ability to measure tiny forces. However, the signals used for imaging consist of voltage differences. Quantitative measurement of forces requires these signals to be converted from voltages into units of force. In practice, this task involves the determination of the cantilever’s stiffness constants and the sensitivities of the optical lever defined in the previous section.

Sensitivities

The value of the normal sensitivity S_n can be determined from a load curve when imaging a hard substrate: its value is given by the slope of the plot of V_{A-B} versus the piezoelectric vertical displacement z_p in the region of the load curve where the cantilever and the substrate are in contact (i.e., $S_n = \Delta V_{A-B} / \Delta z_p$). This procedure may produce erroneous results in the case of soft substrates, for which the normal deformation of the substrate is not negligible compared to the cantilever’s deflection (although in these cases this procedure can be used to measure the stiffness of the surface). The value of S_n depends on the position in which the laser beam falls on the cantilever and thus it has to be measured each time

a cantilever is placed on the AFM or the laser spot is repositioned on the cantilever.

The value of the lateral sensitivity S_l is not so straightforward to obtain as the normal sensitivity. A variety of methods have been published to determine S_l . They all are based on applying a combination of forces that cause a simultaneous torsion and deflection of the cantilever in such a way that the ratio of normal deflection Δz_c to angle of torsion $\Delta\phi$ is known. Their differences arise from the procedure implemented, for example, scanning over substrates with ridges whose slopes form a known angle (Varenberg et al. 2003), acquiring load curves on the slopes of such substrates (Asay and Kim 2006), applying an off-center loading on the cantilever either while scanning a flat substrate (Quintanilla and Goddard 2008), or acquiring a load curve (Feiler et al. 2000). In the latter case, the off-center loading can be achieved by replacing the cantilever's tip with a fiber or a particle in the case of colloidal probes.

Measurement of the lateral sensitivity S_l is complicated by the fact that the description of the optical lever given when presenting the principle of operation is not completely accurate: in many cases the deflection of the cantilever also contributes to the C-D signal. This effect is commonly referred to as "cross-talk" between signals and arises from the fact that a cantilever deflection causes a movement of the laser spot parallel to the line separating the regions C and D of the photodiode only for a unique orientation of the cantilever with respect to the photodiode. Cross-talk between signals appears when the cantilever is placed at an angle with this orientation. Another factor contributing to cross-talk is that the electronic gain applied to the C-D signal is much larger than that applied to the A-B signal to compensate for the fact that the torsion angle of a rectangular or V-shaped cantilever is much smaller than the angular deflection, even if the normal and tangential forces are in the same range. In consequence, even a small cantilever misalignment may allow the deflection to contribute to a significant fraction of the C-D signal.

Cantilever Mechanical Properties

The theoretical expressions for the spring constant of V-shaped and square cantilevers (Sader 2003; Sarid 1994) are:

$$K_n = \frac{Et^3c}{4L^3} \quad \text{square;} \quad (4)$$

$$K_n = \frac{Et^3d}{2L} \left[1 + \frac{4\bar{d}^3}{b^3} \right]^{-1} \quad \text{V - shaped,}$$

where t is the cantilever's thickness, L its length up to the position of the tip, E its material Young's modulus, c is the cantilever's width in the case of rectangular cantilevers and for V-shaped cantilevers b is the distance between the outer edges of the two beams, and \bar{d} is given by:

$$\bar{d} = d \left(1 + \frac{b^2}{4L^2} \right)^{-1/2} \quad (5)$$

where d is the width of the beams forming the V. For a rectangular cantilever the torsional spring constant is given by:

$$K_\phi = \frac{Et^3c}{6(1+\nu)L} \times \left\{ 1 - \frac{\tanh\left[\frac{L}{c}\sqrt{6(1-\nu)}\right]c}{\sqrt{6(1-\nu)}} \frac{c}{L} \right\}^{-1} \quad (6)$$

which means that for rectangular cantilevers of large aspect ratio ($L/c \gg 1$), their torsional stiffness is proportional to their normal stiffness.

As cantilevers are mounted at an angle to the substrate, the normal and horizontal directions on the substrate often do not coincide with the same directions on the plane of the cantilever. The spatial arrangement of cantilever and substrate must be taken into account to calculate the deformation of the cantilever caused by a given set of forces from the substrate.

Guides of commercially supplied cantilevers usually show a typical value of the spring constant K_n . However, since the spring constant K_n depends on the cube of the thickness t , small variations in a cantilever's thickness can cause significant departures from the nominal value. For this reason, it is good experimental practice to calibrate each cantilever individually when measuring forces using the AFM.

Cantilevers can be calibrated by measuring the shift in the resonant frequency after adding known masses (Cleveland et al. 1993), by Doppler vibrometry (Ohler 2007) or measuring the resonant frequency and quality factor of their spectra when they experience forced vibrations (Sader et al. 1999) or are simply excited by thermal noise (Hutter and Bechhoefer 1993).

Surface Characterization

Interaction Potential

As was mentioned in the section above, "Force Spectroscopy: Load Curves," in most cases when the cantilever's tip is approached to the substrate, at some small tip-substrate separation the tip suddenly bends towards the substrate. This "jump into contact" is caused by attractive forces acting between substrate and tip (Burnham and Richard 1989), resulting in an energy potential $U_1(z)$ where z is the

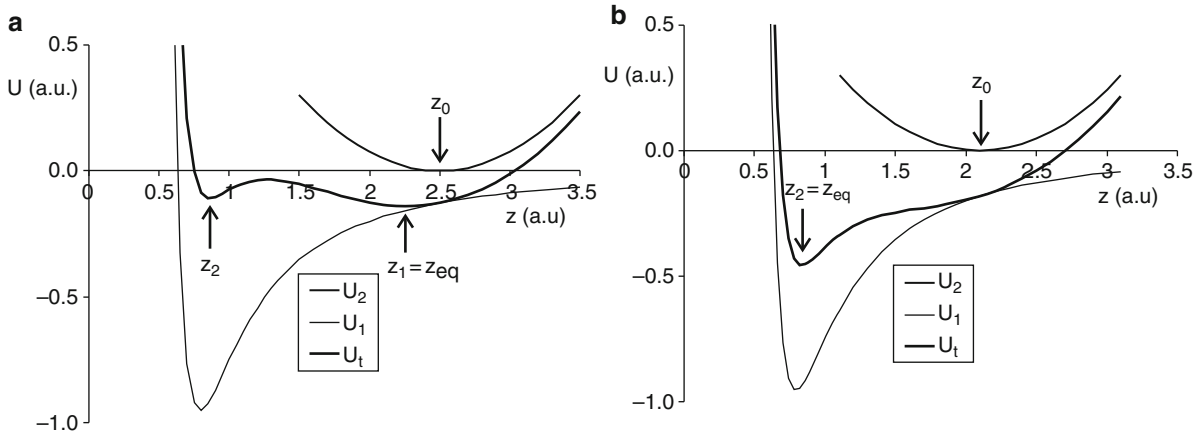
distance between the cantilever's tip and the substrate. The elastic potential energy of the cantilever is $U_2(z) = K_n(z-z_0)^2$, where z_0 is the equilibrium position of the free cantilever and the total energy potential $U_t(z)$ is given by $U_1(z) + U_2(z)$. The shape of $U_t(z)$ depends on the relative strength of the potentials U_1 and U_2 . If the substrate-tip interaction potential U_1 is weak compared with the cantilever's elastic potential U_2 , U_t will have only one minimum for all values of the tip-substrate separation z . The equilibrium position of the cantilever z_{eq} corresponds to the minimum in U_t , which deviates steadily from z_0 until the cantilever makes contact with the substrate. In this case there will be no "jump into contact" and the shape of the tip-substrate interaction potential can be recovered from the shape of the load curve when the tip is close to the surface. However, if the attractive part of the tip-substrate interaction potential U_1 is strong, the total energy potential U_t has two minima at positions z_1 (close to the surface) and z_2 (close to z_0) when the tip-substrate separation is larger than a certain distance z_j and only one minima for $z < z_j$, as shown in the example in Fig. 6. For $z > z_j$ the equilibrium position of the cantilever z_{eq} corresponds to the minimum z_2 even if $U_t(z_1) < U_t(z_2)$ because the cantilever cannot pass the potential barrier between the two minima. At $z = z_j$ the tip will jump to the minimum at z_1 as the minimum z_2 disappears: this instability corresponds to the "jump into contact." In this case, the surface-tip interaction potential can only be probed up to the position of z_j .

In practical terms, contact mode AFM is not the best means to probe the surface interaction potential. Experiments of this type are better performed oscillating

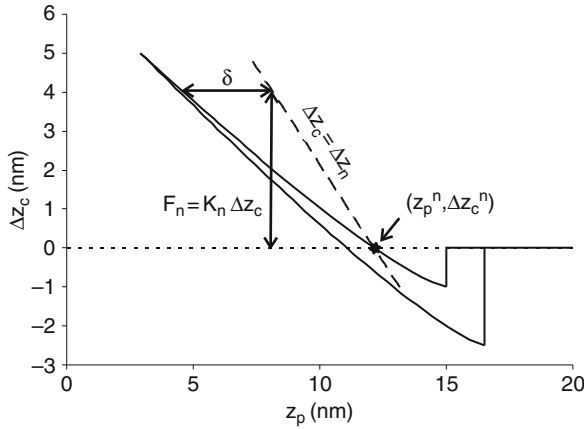
the cantilever inside the potential U_t . Since the cantilever's effective normal stiffness would be $K_n^{eff} = K_n + d^2 U_1 / dz^2$, the difference between K_n^{eff} and K_n causes a shift in the resonant frequency of the cantilever from which $d^2 U_1 / dz^2$ can be recovered. However, the fact that the equilibrium position of the cantilever changes as it approaches the substrate has important consequences for the next section.

Surface Elasticity

When load curves are acquired against a substrate whose elastic compliance is comparable to that of the normal cantilever stiffness, it cannot be assumed that the piezo displacement Δz_p and the cantilever displacement Δz_c are equal when substrate and cantilever's tip are in contact, as it has been done in Fig. 5 and its accompanying discussion. The reason is that the deformation of the substrate is not negligible compared to the cantilever's displacement. In this situation, part of the piezo displacement Δz_p corresponds to the cantilever displacement Δz_c and part of it to the sample deformation δ (Weisenhorn et al. 1993; Bonnell 2001). The sample deformation can be recovered from the load curve if the normal photodiode sensitivity S_n has been measured previously by acquiring a load curve against a hard substrate. If so, the cantilever displacement Δz_c can be calculated from the values of the A-B signal using Eq. 2 to present the load curve in the form of Δz_c values against the piezo position z_p as shown in Fig. 7. The next step is to decide which point in the load curve corresponds to contact without deformation of the substrate: in what follows, this point will be called the neutral point. Note that, according to the previous section, in the



Surface Analysis Using Contact Mode AFM, Fig. 6 Instability in the deflection of the cantilever caused by the evolving shape of the total potential energy of the cantilever $U_t(z) = U_1(z) + U_2(z)$ as the cantilever's tip approaches the substrate. $U_1(z)$: Tip-substrate interaction potential; $U_2(z)$ cantilever's elastic potential. Coordinate z is the distance between substrate and tip



Surface Analysis Using Contact Mode AFM, Fig. 7

Procedure to calculate the surface deformation δ as a function of the load force F_n for the indentation of a soft substrate. The point $(z_p^n, \Delta z_c^n)$ represents the neutral point for which the surface deformation is assumed to be zero. The election of the position of the neutral point requires a model for the tip-substrate interaction

neutral point Δz_c does not necessarily equal zero and it is necessary to have a model for the tip-substrate interaction to give a value for the cantilever deflection Δz_c^n at the neutral point. In Fig. 7, Δz_c^n has been arbitrarily chosen equal to zero for the sake of simplicity. If z_p^n is the piezo position corresponding to the deflection Δz_c^n , then the line $\Delta z_c = \Delta z_c^n + (z_p^n - z_p^0)$ gives the piezo position Δz_c^0 for a cantilever displacement Δz_c if the surface deformation were zero. Figure 7 has been plotted using the convention that a retraction of the cantilever increases the value of the coordinate z_p . Then, the surface deformation is:

$$\delta = z_p^0 - z_p = (\Delta z_c - \Delta z_c^n) - (z_p^n - z_p) \quad (7)$$

while the load force is calculated from Eq. 3. Negative values of the sample deformation are possible depending on the values of Δz_c^n and Δz_p ; they are perfectly reasonable if there is a strong attractive interaction between substrate and tip. If the load curve retraces itself when the tip is retraced, then deformation of the substrate is elastic. In this case, the reduced Young's modulus of the substrate can be obtained from the dependence of the load force F_n with the surface deformation if the form of the cantilever's tip is known. For example, for a conical indenter of half angle α :

$$F_n = \frac{2E^*}{\pi \tan \alpha} \delta^2 \quad (8)$$

where E^* is the reduced Young's modulus of the substrate since the tip is assumed to be perfectly rigid.

Surface Topography

Surface topography analysis of AFM data uses the same mathematical tools as stylus profilometry. The most important parameters used are the root mean square roughness R_q , the arithmetical roughness R_a , the bearing curve, the power spectral density (PSD) of the surface, and its closely related autocorrelation function. The PSD and the autocorrelation function are used to characterize the structure of the surfaces. The bearing ratio is defined as the area of the image above a certain depth measured from the highest pixel in the image and it is useful for estimating the true area of contact between surfaces. The roughness parameters are defined as:

$$R_q = \frac{1}{N} \left[\sum_{i=1}^N (z_i - z_m)^2 \right]^{1/2} \quad (9)$$

$$R_a = \frac{1}{N} \sum_{i=1}^N |z_i - z_m| \quad (10)$$

where z_m is the mean height of the pixels of the surface. The roughness of a surface affects its adhesion, friction, and wear.

Roughness values measured from an image depend on the scan size (Sedin and Rowlen 2001), because when the scan size is increased, irregularities of the surface with longer wavelength are included in the image and contribute to the value of the roughness while the smaller wavelengths are lost due to the increase in the lateral size of each pixel in the image. In short, the lateral size of the pixel acts as the higher spatial cut-off sampling frequency and the image size as the lower cut-off sampling frequency for surface irregularities. The value of the roughness is dominated by the longer wavelengths and therefore, its value increases with scan size until a constant value is reached.

When measuring roughness in small scan size images (pixel size comparable to tip size), it is important to account for the effect of tip broadening due to the finite size of the tip and tip/substrate deformation. For soft solids, it is important to minimize the force exerted by the tip on the sample to limit the deformation. Large substrate deformations or a broadening of the tip, for example, due to tip wear, causes a reduction in the roughness as the tip is unable to trace the surface in the summits between sharp features.

Tip broadening causes a reverse effect in intermediate scan sizes, when the pixel size is not much smaller than the tip size. In this case, as the tip becomes blunter, small

features that would not have appeared in the image if a sharp tip were used cause a reverse image of the tip that contributes to the roughness of the image. The effect is equivalent to a shift from spatial frequencies slightly larger than the high cut-off associated with the pixel size to somewhat smaller frequencies. The extra contribution from frequencies slightly larger than the high cut-off causes an increase in the roughness over its real value for the scan size used, an effect that becomes more important as the tip wears.

Regarding the power spectral densities (PSD), the effect of tip broadening generally does not affect its overall shape. However, the PSD of a surface usually shows a number of peaks at lower frequency, corresponding to the large features in the image whose position may be affected. The lack of change in the general shape of the PSD curve due to tip broadening makes it difficult to detect this effect in PSD curves.

Cross-References

- [Adhesion Hysteresis](#)
- [Adhesive Contact of Elastic Bodies](#)
- [Amontons Laws of Friction](#)
- [Asperities](#)
- [Basic Concepts in Adhesion Science](#)
- [Capillary Force and Surface Wettability](#)
- [Electrostatic Field Effects on Adhesion](#)
- [Friction \(Concepts\)](#)
- [Friction Force Microscopy](#)
- [Stylus Profilometry](#)
- [Surface Free Energy](#)
- [Surface Roughness](#)
- [Van der Waals Forces](#)
- [Work of Adhesion and Work of Cohesion](#)

References

- D.B. Asay, S.H. Kim, Direct force balance method for atomic force microscopy lateral calibration. *Rev. Sci. Instrum.* **77**, art. no. 043903 (2006)
- B. Bhushan, H. Fuchs (eds.), *Applied Scanning Probe Methods*. NanoScience and Technology, vol. I–XIII (Springer-Verlag, Berlin Heidelberg, 2009). ISBN 978-3-540-88823-9
- G. Binnig, C.F. Quate, Ch. Gerber, Atomic force microscope. *Phys. Rev. Lett.* **56**(9), 930–933 (1986)
- D. Bonnell (ed.), *Scanning Probe Microscopy and Spectroscopy: Theory, Techniques and Applications* (Wiley-VCH, New York, 2001). ISBN 0-471-24824-X
- P.C. Braga, D. Ricci (eds.), *Atomic Force Microscopy: Biological Methods and Applications* (Totowa, New Jersey, 2004). ISBN 1-58829-094-8
- N.A. Burnham, J.C. Richard, Measuring the nanomechanical properties and surface forces of materials using an atomic force microscopy. *J. Vac. Sci. Technol.* **7**, 2906–2913 (1989)

- J.P. Cleveland, S. Manne, D. Bocek, P.K. Hansma, A non-destructive method for determining the spring constant of cantilevers for scanning force microscopy. *Rev. Sci. Instrum.* **64**, 403–305 (1993)
- A. Feiler, P. Attard, I. Larsona, A calibration method for lateral forces for use with colloidal probe force microscopy cantilevers. *Rev. Sci. Instrum.* **71**, 2746–2750 (2000)
- J.L. Hutter, J. Bechhoefer, Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* **64**, 1868–1873 (1993)
- B. Ohler, Cantilever spring constant calibration using Doppler vibrometry. *Rev. Sci. Instrum.* **78**, art. no. 063701 (2007)
- M.A.S. Quintanilla, D.T. Goddard, A calibration method for lateral forces for use with colloidal probe force microscopy cantilevers. *Rev. Sci. Instrum.* **79**, art. no. 023701 (2008)
- J.E. Sader, Susceptibility of atomic force microscope cantilevers to lateral forces. *Rev. Sci. Instrum.* **74**, 2438–2443 (2003)
- J.E. Sader, J.W.M. Chon, P. Mulvaney, Calibration of rectangular atomic force cantilevers. *Rev. Sci. Instrum.* **70**, 3967–3969 (1999)
- D. Sarid, *Scanning Force Microscopy* (Oxford University Press, New York, 1994). ISBN 0-19-509204-X
- D.L. Sedin, K.L. Rowlen, Influence of tip size in AFM roughness measurements. *Appl. Surf. Sci.* **182**, 40–48 (2001)
- M. Varenberg, I. Etsion, G. Halperin, An improved wedge calibration method for lateral force in atomic force microscopy. *Rev. Sci. Instrum.* **74**, 3362–3367 (2003)
- A.L. Weisenhorn, M. Khorsandi, S. Kansas, V. Gotz, H.-J. Butt, Deformation and height anomaly of soft surfaces studied with an AFM. *Nanotechnology* **4**, 106–113 (1993)

Surface Analysis Using Dynamic AFM

ALESSANDRO PODESTÀ

Department of Physics and CIMAINA, University of Milano, Molecular Beams and Nanocrystalline Materials Laboratory, Milano, Italy

Synonyms

[Dynamic AFM](#); [Intermittent contact atomic force microscopy](#); [Tapping-mode atomic force microscopy \(TM-AFM\)](#)

Definition

The atomic force microscope (AFM) provides high-resolution, truly three-dimensional imaging capability of interfaces by monitoring the interaction between a nanomechanical probe and the imaged surface. In dynamic AFM (DAFM) the AFM probe is oscillated near its resonance and brought in (intermittent) contact with the surface. The extreme sensitivity of the nanomechanical oscillator to any interaction forces allows exploitation of DAFM for non-invasive high-resolution imaging as well as for mapping interfacial properties at a time.

Scientific Fundamentals

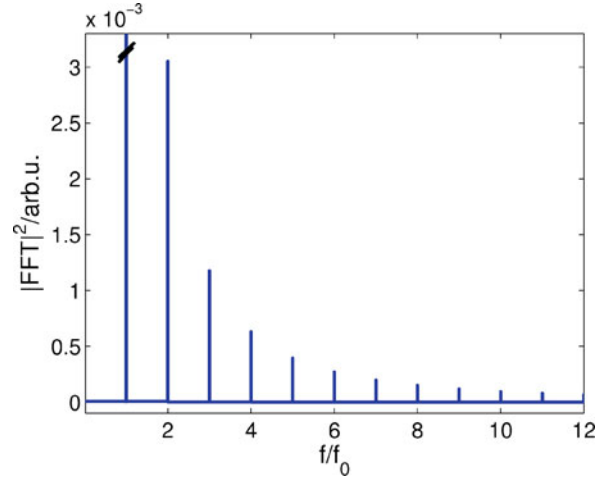
The reader can find in other chapters of this book the description of the basic working principles of the AFM. Here, focus will be on the mechanical model of the vibrating cantilever, and on the basic equations regarding the experimentally accessible quantities. The basic theory underlying the most used DAFM mode, amplitude-modulation (AM)-AFM, will be presented. In AM-AFM the oscillation amplitude is kept fixed at a constant value while the tip is scanned across a surface, in intermittent contact with it. Frequency-modulation-AFM, despite its growing promise as a routine imaging technique, has been up to now restricted mostly to the niche of ultra-high-vacuum investigations of atomically smooth surfaces, with a few relevant exceptions, and will not be considered in the following. For a good review of DAFM techniques, the reader may refer to (Garcia and Perez 2002).

Cantilever Dynamics

A vibrating AFM cantilever, irrespective of its shape, is commonly modeled as a single degree of freedom oscillator, that is as a point effective mass attached to a spring and subject to different forces: viscous damping, a suitable driving force, and a tip-sample interaction. The added mass of the AFM tip located at the free end of the cantilever can be neglected. This simplified model captures with good accuracy the most important details of the cantilever dynamics for most purposes:

$$\frac{d^2 z}{dt^2} + \frac{\omega_0}{Q} \frac{dz}{dt} + \frac{\omega_0^2}{k} [k(z - z_b) - F_{ts}] = 0 \quad (1)$$

where z is the position with respect to the surface (located at $z = 0$), $\omega_0 = 2\pi f_0$, f_0 being the resonance frequency, Q is the quality factor, accounting for the viscous damping and the stiffness of the cantilever, k is the vertical force constant of the cantilever, z_b is the cantilever base position, and F_{ts} is the tip-surface interaction force. Excitation of cantilever vibration can be actuated either by exciting the cantilever base using a piezoelectric actuator (acoustic excitation), in this case $z_b(t) = A_d \sin(\omega t)$, where A_d is the driving amplitude, or applying a driving force F_0 directly on the tip, as in the case of magnetic excitation, in this case z_b is a constant and a term $+ F_0$ appears within the square brackets. Cantilever is typically excited at angular frequency $\omega \approx \omega_0$. The oscillation amplitude A_0 of the single degree of freedom oscillator when it is far from the surface (free oscillation) is peaked around the angular frequency $\omega_r = \omega_0 \sqrt{1 - 1/(2Q^2)}$. The amplitude becomes very small for driving frequencies far away from ω_0 . The tip-surface interaction regime is characterized by an oscillation amplitude $A < A_0$. The cantilever deflection $z - z_b$ is



Surface Analysis Using Dynamic AFM, Fig. 1 Power spectrum of the simulated cantilever deflection shown in Fig. 5a, for an interacting AFM tip. The order of harmonics is reported in the abscissae. Notice the forest of higher harmonics, containing a considerable fraction of the total power

a periodic function with period $T = 2\pi/\omega$, also when the tip is interacting with the surface. The power spectrum (squared modulus of the Fourier transform) of the interacting cantilever contains the higher harmonics of the driving frequency, with frequencies kf_0 , with $k = 2, 3, 4, \dots$. These harmonics contain all the information about the interaction force F_{ts} . Figure 1 shows the power spectrum of the simulated deflection of an interacting cantilever (obtained by numerical integration of the equation of motion 1, see later for details), where the higher harmonics can be clearly observed. The intensity of the higher harmonics monotonically decreases, but still considerable power is stored beyond the tenth harmonic.

If the (root-mean-square) oscillation amplitude is kept constant during imaging, as in standard AM-AFM, the total power is redistributed in the power spectrum from the peak of the fundamental frequency among the higher harmonics. It turns out that the relative intensities of the higher harmonics are very sensitive to changes in the interaction potential, i.e., they are extremely material-sensitive. This aspect will be considered later in the “Key Applications” section.

A more accurate description of cantilever dynamics requires going beyond the single degree of freedom approximation. It is supposed, for sake of clarity, that the cantilever is an elastic beam with rectangular section, and length much larger than its width.

The equation for the dynamic displacement $w(x, t)$ of the cantilever is (Sader 1998).

$$EI \frac{\partial^4 w(x, t)}{\partial x^4} + \mu \frac{\partial^2 w(x, t)}{\partial t^2} = F(x, t) \quad (2)$$

where w is taken with respect to the cantilever base, E is the Young modulus of the cantilever, I is the moment of inertia, μ is the mass per unit length, x is the spatial coordinate along the beam, t the time, and F is the total external force per unit length acting on the beam. Notice that F may have several contributions: the driving force, which excites the cantilever vibration, the viscous damping force from the surrounding medium, and the tip-sample interaction force F_{ts} . The system described by Eq. 2 is characterized by a numerable infinity of flexural eigenmodes, with increasing frequencies f_i with $i = 1, 2, 3, \dots$. For a rectangular homogeneous beam, the ratio f_n/f_1 with $n = 2, 3, 4, \dots$ is approximately 6.3, 17.5, 34.4, etc. The power spectrum of the freely-oscillating cantilever beam consists now of a family of peaks, corresponding to the normal modes. When the AFM interacts with the surface this picture gets more complicated, because the higher harmonics of the driving frequency appear. Some of these higher harmonics ($k = 5-8$ and $k = 16-19$) nearly match the second and the third normal mode frequencies. Beyond those frequencies, the power spectrum carries typically negligible information. The coupling of the higher harmonics with the higher normal modes determines the enhancement of those harmonics with frequencies close to those of higher normal modes. This is a peculiar effect of the continuum elastic cantilever, which cannot be accounted for by the simplified single degree of freedom model, where there is a monotonically decreasing trend of the higher harmonics intensity. The multi-modal description of cantilever dynamics is therefore important in those applications that require the harmonic analysis of the cantilever oscillation in order to get insights on the interaction force F_{ts} .

Tip-Sample Interaction Model

A suitable model for the interaction force F_{ts} includes both a van der Waals attractive contribution for non-contact ($z - a_0 > 0$), and a purely elastic repulsion for contact ($z - a_0 < 0$):

$$F_{ts} = \begin{cases} -\frac{HR}{6z^2} & \text{for } z - a_0 > 0 \\ -\frac{HR}{6a_0^2} + \frac{4}{3}E^*\sqrt{R}(a_0 - z)^{3/2} & \text{for } z - a_0 \leq 0 \end{cases} \quad (3)$$

where a_0 is a suitable interatomic separation, H is the Hamaker constant, R is the tip radius, E^* is the effective Young modulus, taking into account the elasticity of both the tip and the sample. Here the Derjaguin-Muller-Toporov (DMT) model derived in the framework of continuum elasticity theory has been used to account for the (purely) elastic force at contact (Derjaguin et al. 1975). The DMT model is well suited to describe the contact mechanics between stiff bodies with low surface energy and small radii (this case fits well to AFM; when the surface is very soft, and the tip large, the Johnson-Kendall-Roberts (JKR) model can be used instead (Johnson et al. 1971)). The model interaction in Eq. 3 does not contain dissipative terms accounting for the viscoelasticity of the surface, nor does it take explicitly into account capillary interactions (Butt and Kappl 2009), which are always present in ambient conditions. The model can be made more realistic when employed in numerical simulations by adding terms in both non-contact and contact regimes.

Phase Shift

An important measurable quantity of the cantilever dynamics is the phase lag ϕ between the driving force and the AFM tip displacement, which is nearly sinusoidal for sufficiently stiff cantilever and small oscillation amplitudes:

$$z(t) - z_b = A(\omega) \cos(\omega t + \phi) \quad (4)$$

where $A(\omega)$ is the oscillation amplitude of the interacting tip. Remarkably, $\sin(\phi)$ turns out to be a measure of the energy dissipated during each oscillation cycle due to viscous damping and non-conservative interaction forces (Anczykowski et al. 1999):

$$\sin(\phi) = \frac{\omega}{\omega_0} \frac{A(\omega)}{A_0} + \frac{QE_{dis}}{\pi k A_0 A(\omega)} \quad (5)$$

Here A_0 is the free oscillation amplitude, far away from the surface, and E_{dis} is the energy dissipated by non-conservative tip-sample interactions. In AM-AFM, which is the most widely used dynamic mode in standard (non-ultra-high-vacuum) conditions, the oscillation amplitude is constant, and such is the first term in Eq. 5. Consequently, purely elastic interactions ($E_{dis} = 0$) in AM-AFM do not induce any phase shifts.

Key Applications

Phase Imaging

Phase imaging consists in recording the phase lag ϕ between the driving force and the oscillation of the

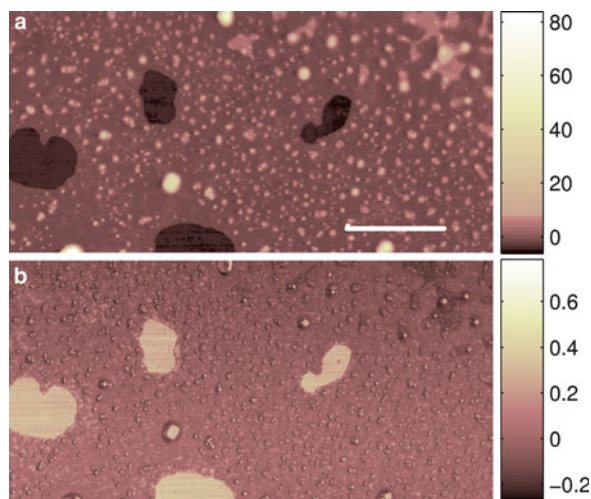
cantilever while the feedback keeps the oscillation amplitude at a fixed setpoint value, typically 10–30% less than the free oscillation amplitude. As detailed in the previous section, $\sin(\phi)$ is related to the energy dissipation due to the tip-sample interaction, this being likely a strongly material-dependent effect. The concurrent acquisition of topographic and phase maps (actually maps of $\sin(\phi)$) can be very useful in highlighting nanoscale surface heterogeneities. At a qualitative level, phase maps provide an immediate feedback on the co-presence, in the investigation area, of regions with different interfacial properties; quantitatively, the phase map can be converted into a dissipated energy map, and using suitable interaction models, this latter can provide some insights on the tip-surface interaction. All the quantities appearing in Eq. 5 are known, or measurable, except the dissipated energy: the force constant k can be calibrated using the thermal noise method (Butt and Jaschke 1995) or the Sader method (Sader 1995); the quality factor Q and the angular frequency $\omega_0/2\pi$ can be obtained by fitting the power spectral density of the thermal fluctuations of the cantilever (Xu and Raman 2007); $A(\omega)$ and A_0 are set by the operator before imaging. Figure 2 shows a pair of topographic and $\sin(\phi)$ maps acquired simultaneously on a lipid bilayer film deposited on a glass coverslip (Indriieri et al. 2008). The phase contrast reveals the presence of the substrate at the bottom of the holes, while it can be noticed that the third incomplete, disordered layer in the upper-

right corner provides a contrast that differs from that of the first bilayer.

Topographic contributions to the phase map must be taken into account when the roughness of the surface is comparable to the tip radius. This is due to the finite response time of the feedback, which cannot keep the tip from getting in more intimate contact (or losing contact with) the surface for a finite time (of the order of fraction of a millisecond); this changes in turn the instantaneous oscillation amplitude and therefore induces a spurious phase shift. This effect can be turned into a benefit when imaging with low forces, because the topography-induced phase shifts highlight fine changes in surface slope and edges, with usually a better contrast than the topographic map.

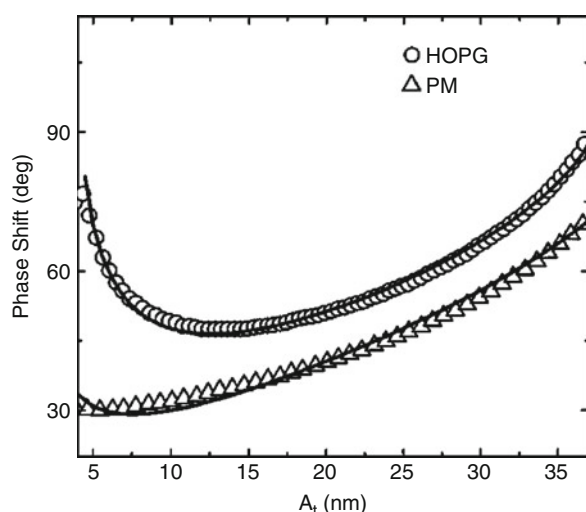
Amplitude Versus Phase Spectroscopy

An interesting application of Eq. 5 is the investigation of the average power (or energy) dissipated per oscillation cycle during the tip-surface interaction via amplitude vs. distance curves (Tamayo and Garcia 1998). Instead of recording the static cantilever deflection as a function of the relative tip-surface distance, which is the base of the classical force vs. distance curves (also called load curves), the cantilever is oscillated at a frequency close to the resonance frequency with a constant driving force (initial free amplitude A_0), and the oscillation amplitude and the phase shift are recorded during approaching-retracting cycles. Plotting $\sin(\phi)$ as a function of A , one obtains the curve of Eq. 5. This curve can be fitted and the parameter E_{dis} is obtained (or alternatively the average power $E_{dis}/(2\pi/\omega)$). It turns out that the power is rather constant irrespective to the amplitude setpoint, such that it can be considered a constant fitting parameter. Remarkably, the dissipation energy E_{dis} calculated from dynamic amplitude vs. phase curves agrees well with that calculated from quasi-static force vs. distance curves, as the area enclosed in the hysteresis of the approaching and retracting branches (Tamayo and Garcia 1998) (see chapter “Surface Analysis Using Contact Mode AFM”). Typical values are 10–500 eV per cycle. Although these values can be far sufficient for breaking covalent bonds, they are actually distributed across an area including several thousands of atoms for a tip radii of 10–100 nm, therefore the energy dissipated per atom is only a small fraction of an eV. This is in agreement with the common observation that for typical tapping conditions no material removal or irreversible damage take place. An example of amplitude vs. phase spectroscopy is shown in Fig. 3. Micrometer-sized patches of purple membrane (soft) has been deposited on graphite (hard), and



Surface Analysis Using Dynamic AFM, Fig. 2 AFM (a) topographic, and (b) $\sin(\phi)$ maps of a phospholipid bilayer deposited on a glass coverslip. Bar is 1 μm . Vertical scale units in (a) are nm (Adapted from (Indriieri et al. 2008))

amplitude vs. phase curves have been measured on both surfaces, according to the methodology described above. Equation 5 was used to fit the experimental data, and values of $E_{dis} = 180$ eV and $E_{dis} = 510$ eV were obtained for the membrane and the graphite, respectively. The goodness of the fit is an a posteriori

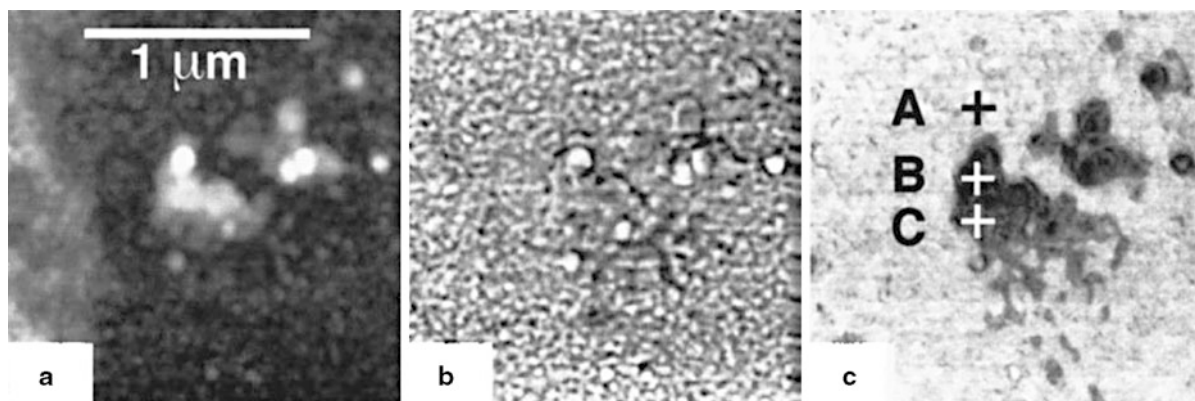


Surface Analysis Using Dynamic AFM, Fig. 3 Phase shift dependence on the tapping amplitude measured during approaching to a *purple* membrane film (*triangles*) and to a graphite surface (*circles*). The *solid lines* represent the fit obtained using Eq. 5. Instrumental data: $f/f_0 = 1$, $A_0 = 43$ nm, $k = 45$ N/m, and $Q = 270$ (Adapted from (Tamayo and Garcia 1998))

confirmation that the assumption of a constant power dissipation rate was correct.

Imaging Using Higher Harmonics

The tip-sample interaction has a strong impact on the power spectrum of the cantilever oscillation. Higher harmonics of the driving frequency show up and couple with the cantilever normal modes, providing a channel for mapping surface heterogeneities with extremely high sensitivity to material properties (Hillenbrand et al. 2000). Using a lock-in amplifier, it is possible to record the amplitude of those higher harmonics that are more strongly enhanced by the coupling with the higher normal modes. If acquired simultaneously to topography, the higher harmonics maps represent a good, actually better, alternative to the phase map described above for highlighting surface heterogeneities with nanometer spatial resolution. The high sensitivity of higher harmonics components to small changes of the tip-sample interaction makes the corresponding maps rich in nano-scale details, often not visible in the topography map. An example of application of higher harmonics imaging is shown in Fig. 4, where the results of higher harmonics imaging on an etched silicon surface are presented. Notice that the phase map is dominated by topography-induced shifts, highlighting the fine corrugation of the silicon surface. The 13th harmonic map shows in turn a strong contrast that is not visible in the phase map and that is likely due to compositional and/or mechanical differences induced by the etching process. AFM controllers have typically one or two built-in auxiliary lock-in amplifiers, therefore higher harmonics imaging is rather accessible to standard users also. Some commercial AFM controllers



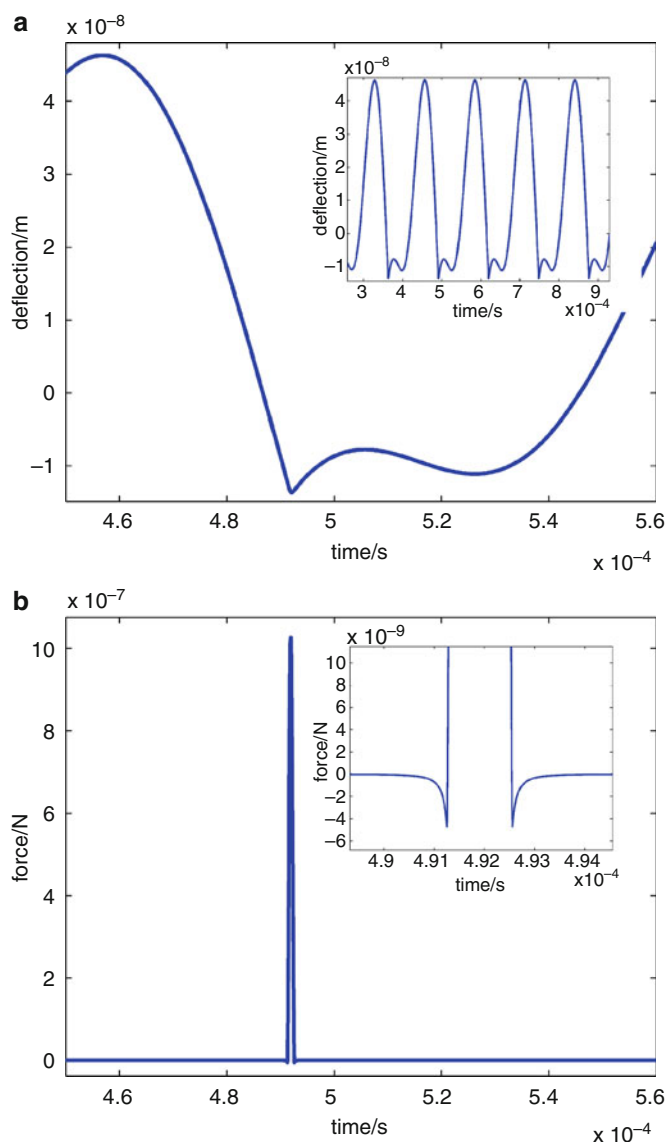
Surface Analysis Using Dynamic AFM, Fig. 4 Images of an etched silicon wafer: (a) topography, (b) phase map, (c) amplitude of the 13th harmonic (Adapted from (Hillenbrand et al. 2000))

have already implemented the higher harmonics imaging as a built-in feature.

Nanomechanical Force Characterization

The analysis of the cantilever deflection signal can also provide quantitative information on the tip-sample force.

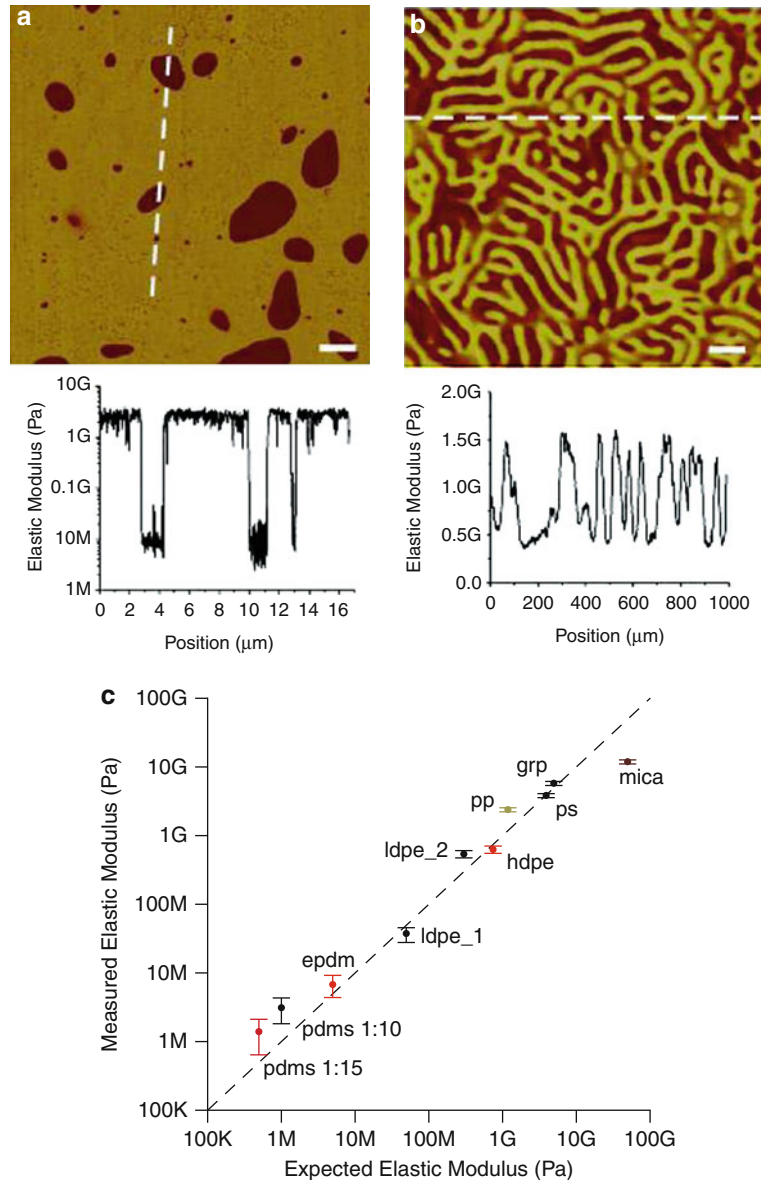
Numerical simulations based on the simplified single degree of freedom model (Eq. 1) allow predicting the effect of anharmonic contributions in the cantilever deflection and to visualize the time profile of the interaction force. Figure 5 shows the simulated deflection $z - z_b$ and force profiles of an acoustically excited cantilever operated in liquid ($Q = 2$, $k = 0.5$ N/m), equipped with



Surface Analysis Using Dynamic AFM, Fig. 5 Numerical simulation of the deflection and force impulse of an AFM cantilever equipped with a silicon tip heavily tapping on a silicon surface in high damping conditions. The simulation is based on the single degree of freedom mechanical oscillator model. (a) Deflection, a single oscillation period (a longer series is shown in the inset); (b) Force impulse (the details of the attractive interaction intervals are shown in the inset)

an oxidized silicon tip heavily tapping on an oxidized silicon surface ($A_0 = 75$ nm, $A = 34$ nm). Other parameters of the simulations are: $E_{tip} = E_{sample} = 169$ GPa, for the Poisson coefficients: $\nu_{tip} = \nu_{sample} = 0.28$, $f_0 = 8$ kHz, $f = 7.8$ kHz, $R = 40$ nm, $H = 6.4 \cdot 10^{-20}$ J.

It can be seen in Fig. 5-a that the effect of tapping on a surface is to distort the sinusoidal oscillation the cantilever has when it is free from any interactions. In the simulation a somewhat extreme case has been considered: a hard tip on a hard surface, at very low setpoint



Surface Analysis Using Dynamic AFM, Fig. 6 Nanomechanical mapping in the tapping mode. (a) Thermoplastic vulcanizate, (b) PMMA-PIB-PMMA block-copolymer film. High-speed force versus distance curves are used to estimate local elastic modulus (logarithmic scale). Below the images are shown numeric values across the sections indicated by the dashed lines. Scale bars are (a) 2 μm, (b) 100 nm. (c) Average values of the effective Young modulus plotted against nominal elastic modulus values of the samples (Adapted from (Sahin and Erina 2008))

amplitude. The force impulse shown in Fig. 5-b has a non-negligible duration (16% of the oscillation period) and the peak force reaches the quite considerable value of 1 μN .

A very recent and interesting development of the harmonic analysis of the dynamic cantilever displacement is the analysis of nanomechanical interaction forces during AFM imaging of surfaces. In the experiments, the time profile of the interaction force could be calculated, according to Eq. 1, as the sum of terms involving the experimentally measured deflection $z - z_b$ and its derivatives. In the Fourier space, the contributions to the cantilever deflection that do not originate from the tip-sample interaction can be easily recognized and removed: only the amplitude stored in the fundamental mode and in the higher harmonics carries information. This strategy is, however, not suitable in general due to the fact that the single degree of freedom model fails in quantitatively describing the relative intensities of oscillation amplitudes of the higher harmonics, because the coupling with the higher normal modes of the beam is not considered.

Following a different approach, the interacting cantilever is treated as a linear system, where an input (the total force acting on it) is transformed into an output (the cantilever deflection measured by the AFM photodiodes) by a linear operator (the transfer function) which can be either measured, or adapted from the theory (Stark et al. 2002; Sahin et al. 2007). The advantage of this approach is that via the transfer function the coupling of the higher harmonics with the normal modes is fully taken into account. Once the transfer function is acquired, the conversion of the deflection signal into a force profile is a matter of matrix operations and Fourier transformations. These transformation can be made fast enough to allow simultaneous force and topography mapping. From each force impulse several parameters can be directly evaluated, like the average interaction force and the peak repulsive and adhesive forces. An interesting application of the force mapping simultaneous to imaging is the possibility of measuring quantitatively some of the parameters characterizing the tip-sample interaction, in particular during the tip-sample contact. According to Eq. 3, once the function $z(t)$ is known (in the contact region, z is the elastic indentation depth), from the mechanical model $F_{ts}(z)$ it is possible to calculate the local (effective) Young modulus of the surface, provided the tip radius is known (Sahin and Erina 2008). Tip radius can be characterized by scanning the AFM tip across a surface containing protruding, very high aspect-ratio features, such that the AFM topography contains inverted tip images. The state-of-the-art of the application of this technique is the real-time nanomechanical mapping of the local Young modulus

across the 1 MPa to 10 GPa range, documented in Fig. 6. Nanomechanical mapping capability is being implemented as a built-in feature in some commercial AFM controllers.

Cross-References

► [Surface Analysis Using Contact Mode AFM](#)

References

- B. Anczykowski, B. Gotsmann, H. Fuchs, J.P. Cleveland, V.B. Elings, How to measure energy dissipation in dynamic mode atomic force microscopy. *Appl. Surf. Sci.* **140**, 376 (1999)
- H.-J. Butt, M. Jaschke, Calculation of thermal noise in atomic force microscopy. *Nanotechnology* **6**, 1 (1995)
- H.-J. Butt, M. Kappl, Normal capillary forces. *Adv. Colloid Interface Sci.* **146**, 48–60 (2009)
- B.V. Derjaguin, V.M. Muller, Y.P. Toporov, Effect of contact deformations on the adhesion of particles. *J. Colloid Interface Sci.* **53**, 314 (1975)
- R. Garcia, R. Perez, Dynamic atomic force microscopy methods. *Surf. Sci. Rep.* **47**, 197–301 (2002)
- R. Hillenbrand, M. Stark, R. Guckenberger, Higher-harmonics generation in tapping-mode atomic force microscopy: insights into the tip-sample interaction. *Appl. Phys. Lett.* **76**, 3478 (2000)
- M. Indrieri, M. Suardi, A. Podestà, E. Ranucci, P. Ferruti, P. Milani, Quantitative investigation by atomic force microscopy of supported phospholipid layers and nanostructures on cholesterol-functionalized glass surfaces. *Langmuir* **24**, 7830 (2008)
- K.L. Johnson, K. Kendall, A.D. Roberts, Surface energy and contact of elastic solids. *Proc. R. Soc. Lond. Ser. A* **324**, 301 (1971)
- J.E. Sader, Method for the calibration of atomic force microscope cantilevers. *Rev. Sci. Instrum.* **66**, 3789 (1995)
- J.E. Sader, Frequency response of cantilever beams immersed in viscous fluids with applications to the atomic force microscope. *J. Appl. Phys.* **84**(1), 64 (1998)
- O. Sahin, N. Erina, High-resolution and large dynamic range nanomechanical mapping in tapping-mode atomic force microscopy. *Nanotechnology* **19**, 445717 (2008)
- O. Sahin, S. Magonov, C. Su, C.F. Quate, O. Solgaard, An atomic force microscope tip designed to measure time-varying nanomechanical forces. *Nat. Nanotechnol.* **2**, 507 (2007)
- M. Stark, R.W. Stark, W.M. Heckl, R. Guckenberger, Inverting dynamic force microscopy: from signals to time-resolved interaction forces. *Proc. Natl. Acad. Sci.* **99**, 8473 (2002)
- J. Tamayo, R. Garcia, Relationship between phase shift and energy dissipation in tapping-mode scanning force microscopy. *Appl. Phys. Lett.* **73**, 2926 (1998)
- X. Xu, A. Raman, Comparative dynamics of magnetically, acoustically, and Brownian motion driven microcantilevers in liquids. *J. Appl. Phys.* **102**, 034303 (2007)

Surface and Bonding

► [Bonding at Surfaces/Interfaces](#)

Surface Characterization and Description

HORST BODSCHWINNA¹, JÖRG SEEWIG²

¹Institut of Measurement and Automatic Control, Leibniz University Hannover, Hanover, Germany

²Lehrstuhl für Messtechnik & Sensorik, Technische Universität Kaiserslautern, Kaiserslautern, Germany

Synonyms

Roughness; Roughness characterization; Roughness measurement

Definition

To a great extent, the quality of a workpiece surface determines the quality of an industrial product. This quality is ensured by a description of the surface properties that is based on the workpiece's intended use and that specifies the limits of the parameters in the drawing. Hence, surface characterization and description are fundamental tasks in surface metrology.

Scientific Fundamentals

The Aim of Surface Characterization and Description

In following the specifications of a surface, the manufacturing steps in surface processing must be coordinated with economy in mind. The functional characteristics of the workpiece surface are determined by (1) macro- and micro-geometrical properties and (2) physical and chemical properties of the near-surface layer. Von Weingraber (Weingraber and Abou-Aly 1989) separates the near-surface layer into an outer and inner boundary layer (Fig. 1).

The outer boundary layer, with its thickness in atomic dimensions (Michaelis 1962), determines the course of many physical and chemical processes due to its free surface energy. There, adsorption and adhesion processes can have a great impact on the friction and abrasion of unlubricated, slide-stressed metal surfaces.

In contrast, the inner boundary layer is the zone in which the material structure is changed by machining. Because in chip producing machining the properties of both the inner and outer boundary layer are affected, Brinksmeier (Brinksmeier 1990) unites both boundary layers under the broader term *surface zone*. The properties of the surface zone incorporate geometrical as well as material characteristics, where the latter result from the

material composition and the production process. The load capacity due to strength, hardness, deformation resistance, and crack sensitivity of the surface zone is affected by residual stress, for instance, from heat treatment or surface machining. In mechanical engineering, these is especially true in the functional performance of tribological contact areas such as dry or lubricated sliding, rolling, or marvering surfaces, and static or translational/rotating sliding sealing faces.

Regarding the functional performance of technical surfaces, in addition to the geometrical characteristics of the workpiece's surface, it is necessary to closely examine the properties of the material composition in the surface zone. Both are included in the term *surface integrity* in US standards (ANSI B211.1-1986).

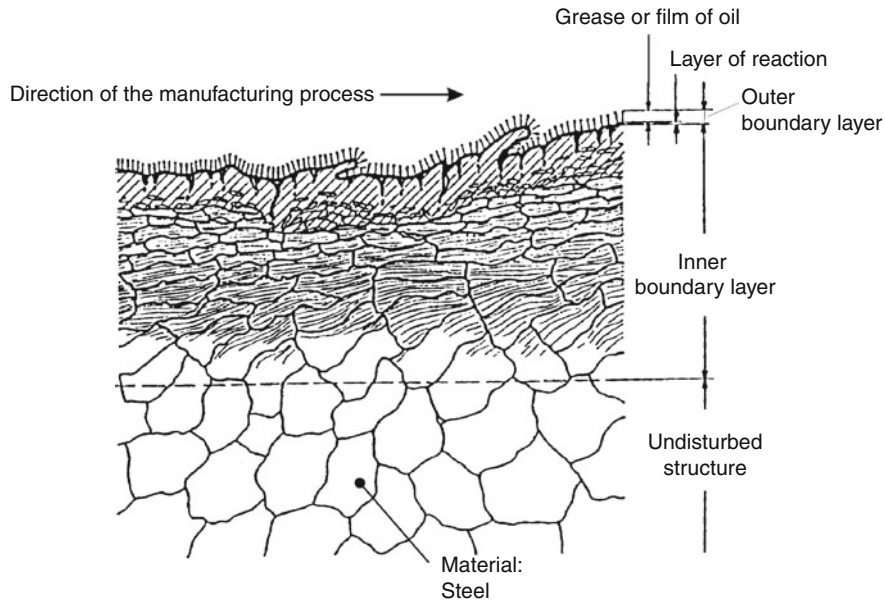
Data Preprocessing for Separation of Form Deviation, Waviness, and Roughness

The surface profile consists of a superposition of the proportion of form and form deviation as well as of waviness and roughness in the measuring section. Figure 2 clarifies this with an example of a nominal straight surface profile where the proportion of long waves can solely be interpreted as the shape deviation.

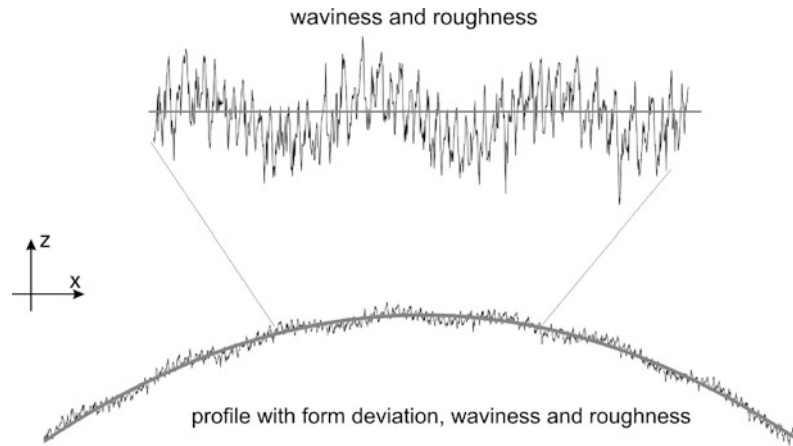
To improve the quality of finishing processes and thus product quality, a functional and production-orientated analysis of the surface profile is compulsory. Bodschiwinna and Hillman (Bodschiwinna and Hillman 1992) demand the separation of form deviations, roughness, waviness, and form for surface characterization because (1) they have different causes in manufacturing and (2) they influence the functional capability of the workpiece differently.

Roughness is determined by the actual condition of the machining process. Reasons for waviness are run-out or form errors, for example, of milling tools, and, most frequently, vibrations (self- or separately induced). Form deviations can be traced back to deviations of the drive unit from the ideal machine tool or to elastic deformation in the system machine-tool and tool-workpiece. Often, however, they are based on deformation due to hardening or release of internal stress during machining, and require an in-depth analysis of the complete manufacturing sequence.

In dealing with the high functional requirements of surfaces, form deviations and waviness must be kept within narrow limits. Only then can a roughness structure based on the intended use be achieved. This applies, in particular, to the contact surfaces in mechanical engineering like static or moving sealing faces, dry and lubricated sliding faces, and rolling and pitch surfaces.



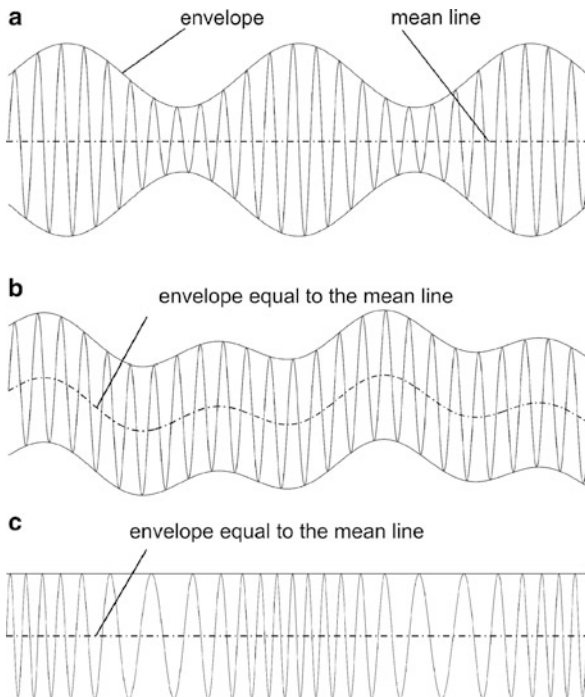
Surface Characterization and Description, Fig. 1 Outer and inner boundary layer of a manufactured workpiece (Weingraber and Abou-Aly 1989)



Surface Characterization and Description, Fig. 2 Superposition of form deviation, waviness, and roughness

In the past, to separate form deviation, waviness, and roughness, different reference line systems have been developed, the effectiveness of which is often discussed using the functional case “contact surface.” Whitehouse (Whitehouse 1994) illustrates the principally different approaches of a peak-orientated reference line (envelope) and a center line reference line with the example of different waveforms of waviness and roughness (Fig. 3).

Von Weingraber’s E system and the MOTIF system used by the French automotive industry determine the profile of waviness (the form profile results from the envelope of the waviness profile) with the help of the profile peaks. These systems are particularly suitable for rating the surface’s mechanical strength from the material-free point of view. The profile of waviness is also the reference line for calculating of roughness parameters. Figure 3 show that by using a peak-orientated reference



Surface Characterization and Description, Fig. 3 Waviness with different waveforms. (a) Amplitude-modulated roughness signal. (b) Additional superposition of waviness and roughness. (c) Wavelength or system phase-modulated roughness signal

line, even amplitude-modulated waveforms (which seldom occur) can be evaluated properly.

In surface measurement, the center line system was established initially on the basis of a 2RC-filter; since 1996, it has been based on a phase-corrected filter with a Gaussian weighting function, according to ISO 11562 (ISO 11562 Geometrical Product Specifications (GPS) 1996). Advantages of the center line system are:

- Signal-theoretical descriptiveness of the transfer behavior of this reference line system (filter process).
- Simple instrumentation requirements, especially with analogue instruments of the past.
- Lower sensitivity of the reference line, especially with single projecting peaks, compared with the E-system.

Due to increasing quality requirements of functional surfaces and improved precision machining methods, the reference line system has to accomplish not only a defined separation of roughness, waviness, form, and form deviations, but also a distortion-free

restitution of the profile's single components, especially of the roughness profile.

With the introduction of the phase-corrected Gaussian filter, according to ISO 11562 (ISO 11562 Geometrical Product Specifications (GPS) 1996), a clear improvement can be achieved by avoiding the phase shift between sinusoidal signal components that occurs due to the application of the causal 2RC filter. This improvement is insufficient, however, if the surface fine finish of mechanically highly stressed contact surfaces is evaluated considering the function and, where applicable, improved by changes to the precision machining process.

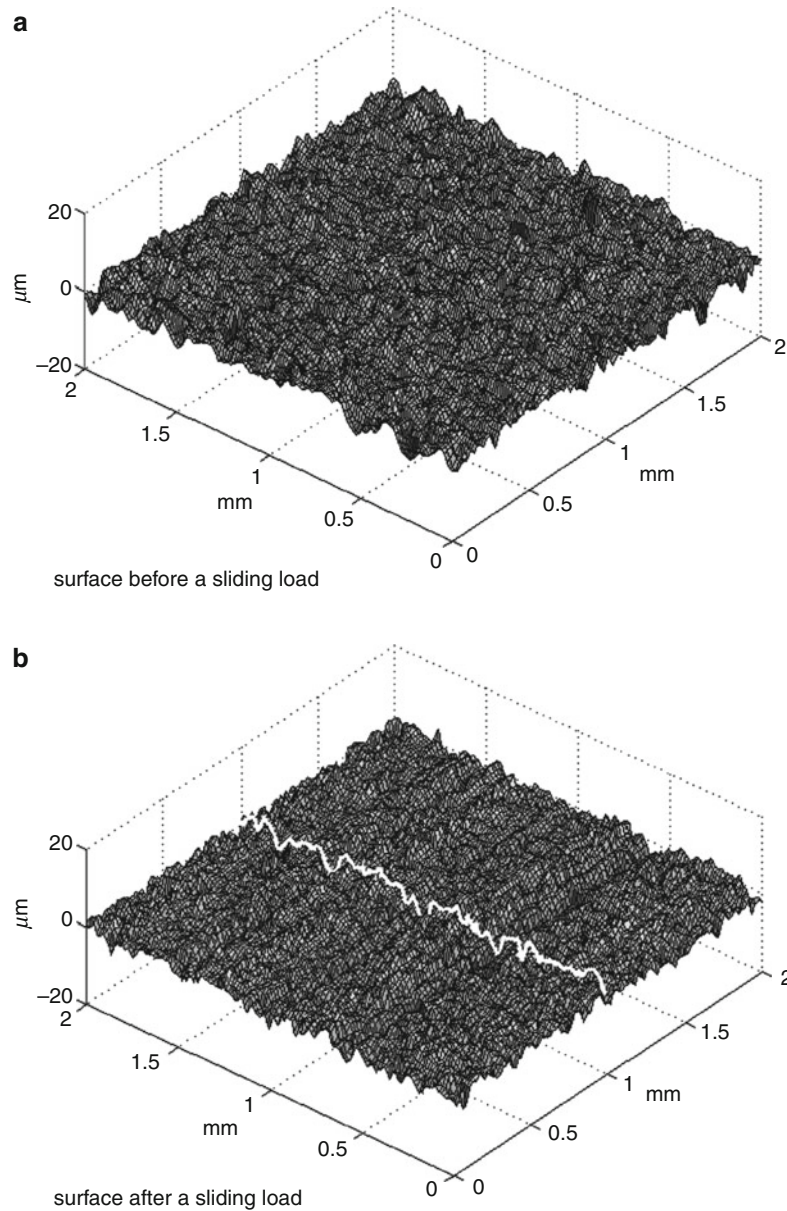
Surfaces of this kind are partly plateau-shaped by multi-stage precise machining processes. The reference line for evaluation of the surface's fine finish is, in this case, better not as a center line, but as one that follows the surface's plateau. The robust phase-corrected Gaussian filter developed by Seewig (Seewig 2005) offers excellent properties for this surface characteristic.

Area of Application of Parameters and Characteristic Functions

The characterization of technical surfaces is subject to the following trends:

- In industrial praxis, profile characteristics are almost exclusively used. Here, predominantly perpendicular measurements, like the average distance between the highest peak and lowest valley, R_z , and the arithmetic average of the roughness profile, R_a , are made.
- In scientific research, sophisticated signal analysis is carried out by application of characteristic functions from the field of signal theory, with stochastic processes like the amplitude density function, the auto-correlation function, and the spectral power density function.

The functional behavior of contact surfaces, such as fit faces and static, translational, and rotational moved sealing faces, depends on the roughness height and on the material distribution in the roughness profile. Whitehouse (Whitehouse 1978) characterizes these properties by parameters of the beta function, which he introduced to approximate the amplitude density function of the roughness. Bodschiwinna (Bodschiwinna 1988) uses the cumulative frequency function of the roughness' amplitudes, which is approximated by a three-straight-line model. This model, which allows for a partition of the overall roughness depth into functionally different depth ranges, is considerably easier to interpret compared with



Surface Characterization and Description, Fig. 4 Change in isotropic character of a blasted surface caused by sliding load. (a) Before sliding load. (b) After sliding load

the beta function. This technique is standardized today in ISO 13565, part 2 (ISO 13565-2 Geometrical product specifications (GPS) 1996), and is used worldwide, particularly in the automotive industry.

In the development of technical surfaces, the trend is toward three-dimensional surface analysis. This

methods allows for considerably better descriptions of surface details and their configuration on the face. According to Stout et al. (1993), 3D evaluation is, in many cases, executed using methods from the stochastic system theory as they are expanded to describe spatial structures.

Key Applications

Topology of Surfaces from Different Manufacturing Methods and Functional Requirements

Figures 4, 5, 6, and 7 show a representative overview of topologies of different manufacturing methods. In the following, guidelines are developed for achieving satisfactory surface descriptions based on surface character and functional requirements. In addition to the 3D figure, a two-dimensional section is plotted in each case to explain the more or less limited informational value of the profile.

Figure 4 displays the 3D surface structure of a ground surface with a one-dimensional directional surface finish. In this case, the 2D roughness profile, which was recorded orthogonal to the direction of the grooves, contains a sufficient amount of information to characterize this surface. For parts with minor functional requirements in surface finish, characteristics like R_z or R_a are usually measured in order to keep surface roughness within specified limits. If the shape of the roughness profile (pointed or round ridged) is to be identified, the ratio of the maximum profile peak height, R_p , to the maximum height of the profile, R_z , is taken.

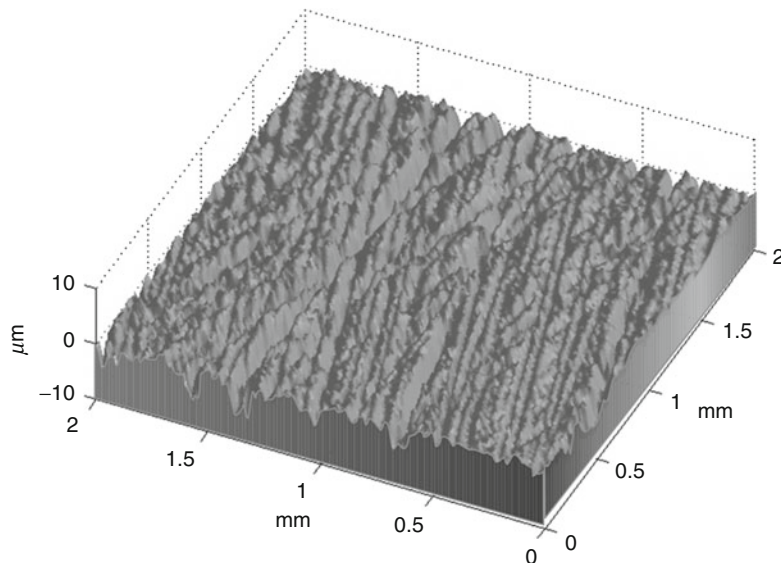
The spatial frequencies and the related wavelengths can be calculated using frequency analysis or the autocorrelation function.

While the new condition of this surface, caused by the random process blasting, can be called isotropic, it shows superposed grooves after the sliding load. Due to this

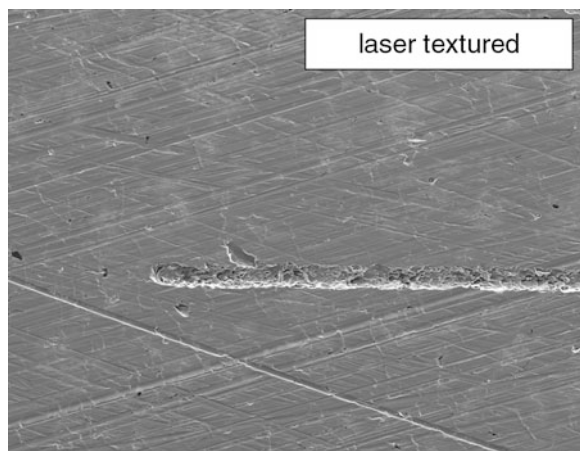
isotropic character, the surface can now be described in any direction in 2D. After the sliding load, the surface obtains an additional directional structure of wear grooves in the direction of displacement (Fig. 3b). They are clearly visible in the topographical overview and can be recorded in the profile. In addition to straightforward characteristics like R_z and R_a , the frequency analysis or the autocorrelation function of a profile's sectional view in various directions – or even laminar – offer a highly informational value for a tribological rating of the surface.

Figure 5 illustrates the surface structure of a honed cylinder running surface. Because the honing angle is known from the manufacturing process, a quality check is possible by 2D section. However, it is insufficient to characterize the surface only by values such as R_z or R_a . This is because functional tribological requirements demand high mechanical strength and thus a plateau-like surface and adequate groove volume to hold oil. Here, parameters like R_{pk} , R_k , and R_{vk} , which present the material distribution in the profile and are deduced from the Abbott curve, have proven their usefulness. The topographical Abbott curve, which was determined in 3D, offers even greater informational value. It allows for evaluation of material-filled or material-free spaces in different sections in the surface topology.

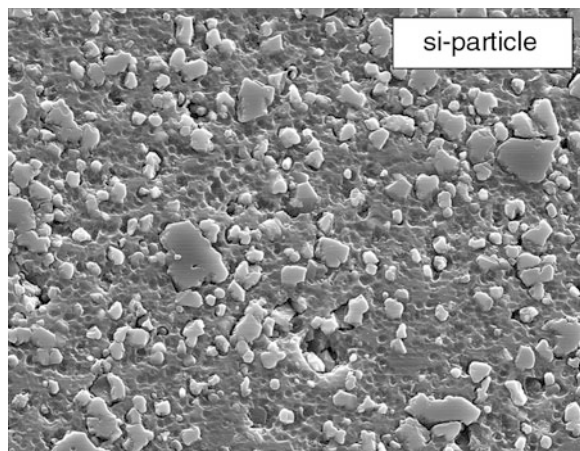
Figure 6 visualizes the structure of a surface that has been laser-textured in order to improve its tribological properties, i.e., to hold lubricating oil. Here, the surface must first be characterized prior to laser patterning. Subsequently, it is necessary to evaluate statistically from the



Surface Characterization and Description, Fig. 5 Surface structure of a honed cylinder running surface



Surface Characterization and Description, Fig. 6 SEM picture of a laser-textured surface



Surface Characterization and Description, Fig. 7 Cylinder running surface with *diamond-honed* Si crystals incorporated into the aluminium

3D measured topology the functional relevance of the laser cups' size, depth, and volume, as well as their abundance and distribution on the surface.

Figure 7 shows the topography of a cylinder running surface of an aluminium cylinder block where silicon crystals cast into the running tread act as tribological micro-contact surfaces. The surrounding aluminium-matrix create micro-cavities for absorption of lubricant oil. Practically, such a surface structure can be measured in 3D only by image processing microscopes and must be evaluated by special software. The aim is to characterize the functionally relevant micro-geometrical properties of the particles and vacancies distributed on the surface.

In summary, rigorous measurement and evaluation is mandatory for topographical measurements in prototype manufacturing, when developing new surfaces with inserted essential micro-details. For quality control in production, it is possible to obtain partial information by profile measurements, which, in combination with previous knowledge about the expected surface character, can be sufficient for this purpose.

Cross-References

- [Surface Statistics and Probability Density Function](#)
- [Surface Synthesis Based on Surface Statistics](#)
- [Surface Variation in Tribological Processes](#)
- [Tribology](#)

References

- ANSI B211.1-1986, American National Standard on surface integrity. Hrsg, SME, Dearborn (1986)
- H. Bodschinna, Funktionsgerechte Rauheitsmessung. Technische Rundschau **28**, 36–40 (1988)
- H. Bodschinna, W. Hillmann, *Oberflächenmeßtechnik mit Tastschnittgeräten in der industriellen Praxis*, Deutsches Institut für Normung e.V. (Beuth-Berlag, Berlin, Köln, 1992), 173 S, ISBN 3-410-12657-0
- E. Brinksmeier, *Prozeß- und Werkstückqualität in der Feinbearbeitung* (Habilitationsschrift, Universität Hannover, 1990), 252 S
- ISO 11562 Geometrical Product Specifications (GPS), Surface texture: profile method – metrological characteristics of phase correct filters (1996)
- ISO 13565-2 Geometrical product specifications (GPS), Surface texture: profile method – Surfaces having stratified functional properties – part 2: height characterization using the linear material ratio curve (1996)
- A.S. Michaelis, Fundamentals of surface chemistry and surface physics. ASTM Special Technical Publication No. 340, in *Symposium on Properties of Surfaces*, Los Angeles, 1962, S. 3–23
- J. Seewig, Linear and robust Gaussian regression filters. J. Phys. Conf. Ser. **13**(1), 254–257 (2005)
- K.J. Stout, P.J. Sullivan, W.P. Dong, E. Mainsah, N. Luo, T. Mathia, H. Zahouani, The development of methods for the characterization of roughness in three dimensions. Commission of the European Communities, Publication No. EUR 15178 EN, 1993, ISBN 0-70441313-2
- H.V. Weingraber, M. Abou-Aly, *Handbuch technische Oberflächen*. Typologie, Messungen und Gebrauchsverhalten. (Vieweg-Verlag, Braunschweig, 1989), 448 S, ISBN 3-528-06318-1
- D.J. Whitehouse, Beta functions for surface typology. Ann. CIRP **27**, 491–497 (1978)
- D.J. Whitehouse, *Handbook of Surface Metrology* (Institute of Physics, Bristol, 1994), p. 988 S. ISBN 0-7503-0039-6

Surface Chemical Analysis Using XPS

- [X-Ray Photoelectron Spectroscopy \(XPS\)](#)

Surface Deformation Calculation for EHL

WEN-ZHONG WANG

School of Mechanical Engineering, Beijing Institute of Technology, Beijing, People's Republic of China

Synonyms

Surface displacement calculation

Definition

The surface deformation calculation for elastohydrodynamic lubrication determines the normal surface displacement due to the pressure distribution generated in the lubricated contact areas.

Scientific Fundamentals

The calculation of normal surface deformation plays a key role in numerical solutions of elastohydrodynamic lubrication (EHL) contacts and consumes a great amount of computer time. The objective is to evaluate the normal surface displacement caused by a distributed load. It is usually assumed that the surface is supported by elastic half-space. For the normal deformation caused by forces in several special forms, such as a point load, a uniformly distributed pressure, or a Hertzian pressure, there are accurate closed-form solutions. However, for a general form of pressure distribution it is impossible to obtain a closed-form solution; thus, numerical techniques have to be employed. In a conventional way, the normal deformation can be numerically calculated by constructing a matrix of influence coefficients and conducting a direct summation based on linear elasticity theory. To obtain influence coefficients, the pressure distribution has to be interpolated on each subdomain according to the nodal pressure values. The interpolation functions include those in the form of a constant, a piecewise biquadratic polynomial, or a bilinear polynomial. For a line contact problem, the direct summation demands as many as N^2 multiplications, where N is the grid number. The number of multiplications is even greater for three-dimensional problems. Such a large amount of computation work greatly limits the efficiency of numerical solvers.

During the past three decades, many efforts have been made to evaluate the surface deformation more efficiently and with less computer storage. A method, known as multilevel multi-integration (MLMI), which is orders of magnitude faster than the conventional methods, was developed by Lubrecht and Ioannides (1991). The method

has proved to be very efficient in saving CPU time, though it is a complicated procedure in programming.

Noting the nature of convolution in determining surface deformation, the fast Fourier transform (FFT) technique has been applied in recent years to the calculations. The FFT approach can give exact results if the surfaces in contact and the pressure can be adequately described by periodic functions. For concentrated contact problems, however, the application of FFT will create so-called periodicity errors. Previous studies in this field tried to reduce the periodicity errors, mostly by extending the computation domain and by zero padding of pressure. However, the increase in the computation domain will certainly ruin the efficiency of the method. Recently, an improved FFT-based method (DC-FFT) (Liu et al. 2000) was proposed resulting from the discrete convolution theorem, which successfully overcomes the periodicity errors. The computation speed is very preferable (Wang et al. 2003). In the following, several numerical algorithms involved in computations of normal surface deformation are presented briefly.

Consider a distributed pressure acting on an elastic half-space, and let the pressure distribution and the normal surface deformation be denoted by $p(x)$ and $v(x)$ for line contacts, or by $p(x,y)$ and $v(x,y)$ for point contacts, respectively. According to the theory of contact mechanics (Johnson 1985), the normal surface deformation $v(x)$ or $v(x,y)$ on one surface caused by a distributed pressure may be written in the forms of

$$v(x) = -\frac{2(1-\nu^2)}{\pi E} \int \ln|x-s| p(s) ds \quad (1)$$

for line loading and

$$v(x,y) = \frac{1-\nu^2}{\pi E} \iint \frac{p(\xi,\eta)}{\sqrt{(x-\xi)^2 + (y-\eta)^2}} d\xi d\eta \quad (2)$$

for point contacts.

In numerical analysis, both functions of normal surface deformation and pressure distribution have to be discretized in a space domain over N grid points for a line load, or $N \times M$ grid points for a two-dimensional distributed load. As an example, the deformation for line loading can be rewritten in discrete form as follows:

$$v(x_i) = -\frac{2(1-\nu^2)}{\pi E} \sum_{j=0}^{N-1} K(x_i-x_j) p(x_j) \quad (3)$$

where $K(x_i-x_j)$, or simply denoted as K_j^i , is known as the influence coefficient that refers to the normal deformation at point x_i due to a unit load acting on a position x_j . It can

be seen from (3) that the calculation of normal surface deformation includes two steps:

1. Determine the influence coefficients K_j^i .
2. Calculate the multi-summation.

For simplicity, the following discussion will be limited to the line contact problems only, and it can be easily extended to the point contact problems.

Determination of Influence Coefficients

The influence coefficient (IC) K_j^i has been interpreted physically as the deformation at point x_i due to unit point load acting on x_j . For the distributed load, K_j^i can be obtained through calculating the deformation at x_p caused by the pressure distributed over a small area around the position x_j :

$$K_j^i = \int_{x_j - \Delta x/2}^{x_j + \Delta x/2} h(x_i - \xi) n(\xi) d\xi \quad (4)$$

where Δx denotes the dimension of a element in discrete grid, $n(x)$ is an interpolation function to approximate the pressure distribution within the element, and $h(x)$ is known as the response or Green's function, which can be written in the form of $\ln(x)$ for the line contact problems. For calculation of normal deformation on a half-space of homogeneous materials, K_j^i depends only on the distance between point x_i and x_j . When x_i is fixed on the coordinate origin, the distance between x_i and x_j is $(x_j - x_i) = x_p$ so that (4) may be rewritten as

$$K_j = \int_{x_j - \Delta x/2}^{x_j + \Delta x/2} h(\xi) n(\xi) d\xi \quad (5)$$

where influence coefficients K_j denote the deformation at the origin due to a load acting on x_j .

Different interpolation functions can be used for determining the influence coefficients, as summarized briefly in the following.

Green' Function-Based Scheme

When $n(x)$ in (5) is taken as a unit constant, and the Green's function $h(x)$ is assumed to be invariant within a element, approximated by $h(x_j)$, influence coefficient K_j can be simply determined as:

$$K_j = \Delta x \times h(x_j) \quad (6)$$

It is noticed that the Green's function $h(x)$ has one singularity when $x_j = 0$, but it can be eliminated by an integration over the element around the point $x_j = 0$. A computation experiment shows that expression (6) may result in a significant numerical error when coarser grids are employed.

Constant Function-Based Scheme

If $n(x)$ in (5) is taken as a constant function on the element surrounding point x_p , the influence coefficients can be gotten by performing the following integration

$$K_j = \int_{x_j - \Delta x/2}^{x_j + \Delta x/2} h(\xi) d\xi \quad (7)$$

which leads to an analytical formula

$$K_j = x_p \ln \left(\frac{2x_p}{\Delta x} \right)^2 - x_m \ln \left(\frac{2x_m}{\Delta x} \right)^2 \quad (8)$$

with $x_p = x_j + \Delta x/2$, $x_m = x_j - \Delta x/2$

For two-dimensional problems, readers can refer to reference (Venner and Lubrecht 2000).

Linear Interpolation Based Scheme

If a linear interpolation function $n(x)$ is applied to approximate the pressure distribution within the element, the influence coefficients can be obtained by performing the integration in (5), which results in

$$K_j = x_j \frac{x_{j-1} + x_{j+1}}{\Delta x} F(x_j) - \frac{x_{j-1}^2}{\Delta x} F(x_{j-1}) - \frac{x_{j+1}^2}{\Delta x} F(x_{j+1}) \quad (9)$$

with $F(x) = 2\ln|x|$, and $F(0) = 0$

For two-dimensional problems, if a bilinear interpolation function is employed, the influence coefficients can be computed likewise in analytical form (Ai 1993).

One may adopt higher-order polynomials, such as a biquadratic polynomial, but the methods are conceptually the same. Theoretically, the numerical accuracy corresponding to higher-order interpolations is expected to be improved, but computation practices show this is not always true. As a matter of fact, when the grid becomes very fine, there is little difference between the results from different orders of interpolation.

Calculation of Multi-summation

Having the influence coefficients obtained, the normal surface deformations can be obtained from the multi-summation as described in (3). The computation may be implemented using different numerical approaches, including direct summation (DS), multi-level multi-integration (MLMI), and DC-FFT based methods, which will be briefly described in this section.

Direct Summation (DS)

A direct computation of (3) may reach accuracy up to the level of discrete error, but this needs N^2 multiplications plus $(N-1)^2$ additions. For two-dimensional problem,

it needs $N^2 \times M^2$ multiplications and $(N-1)^2 \times (M-1)^2$ additions. The computational work will be enormous for very large grid numbers, so a main concern is how to get the results within a reasonable CPU time. Multi-level multi-integration (MLMI) and the discrete convolution and FFT-based method (DC-FFT) are two preferential candidates that can meet the demands for accuracy and efficiency.

DC-FFT Method

Equation (3) is in fact a discrete linear convolution whose calculation can be speeded up by applying the discrete Fourier transform (DFT), but periodic errors may result from two sources:

- (a) Green's function $h(x)$ was truncated at the boundary of computation domain. In the reality, however, function $h(x)$ may possess non-zero values beyond the computation domain. The truncation of $h(x)$ will therefore result in errors in the convolution. The error is expected to be minimal at the center of computation domain but increases at the positions close to the border of the domain.
- (b) When the DFT was employed to calculate the convolution, it means all signals and output were converted into periodic functions. This, more or less but inevitably, will bring about the periodic errors.

To use the DFT properly for evaluating normal surface deformation, the linear convolution in (3) has to be transformed to the circular convolution as proposed by (Liu et al. 2000). This requires a pretreatment for the influence coefficient $\{K_j\}$ and pressure $\{p_j\}$ so that the convolution theorem for circle convolution can be applied. The pretreatment can be performed in two steps:

1. Extend the dimension of $\{p_j\}$ by adding an appropriate number of zeros to the array;
2. Extend the dimension of $\{K_j\}$ to the same length as $\{p_j\}$ and then wrap-around the series $\{K_j\}$.

In summary, when DC-FFT algorithm is applied, the calculation of normal surface deformation within a region $[x_0, x_e]$ can be implemented in following procedure:

1. Discretize the influence coefficient $K(x)$ into a series $\{K_j\}_{2N}$ of size $2N$ in a region whose sides are twice as that of the computation domain.
2. Reorder the indexes of $\{K_j\}_{2N}$. The rule is to copy the components K_1, \dots, K_N of $\{K_j\}_{2N}$, into a new series $\{KN_j\}_{2N}$ as KN_{N+1}, \dots, KN_{2N} , and copy K_{N+1}, \dots, K_{2N} into the series $\{KN_j\}_{2N}$ as KN_1, \dots, KN_N

3. Discretize the pressure function $p(x)$ into a series $\{p_i\}_N$ of size N in the computation domain, and then extend the length of the series $\{p_i\}_N$ into a new series $\{pN_i\}_{2N}$ through zero padding.
4. Apply FFT to the series $\{pN_i\}_{2N}$ and $\{KN_j\}_{2N}$, and the results are denoted as $\{\hat{pN}_i\}_{2N}$ and $\{\hat{KN}_i\}_{2N}$.
5. Compute the component-wise product of $\{\hat{pN}_i\}_{2N}$ and $\{\hat{KN}_i\}_{2N}$, resulting in $\{\hat{T}_i\}_{2N}$.
6. Perform the inverse FFT to $\{\hat{T}_i\}_{2N}$, which gives a new series $\{T_i\}_{2N}$. The desired deformation $\{v_i\}_N$ in the region $[x_0, x_e]$ is obtained by setting $v_i = T_i$, $i = 1, \dots, N$.

The total complexity of the numerical analysis is $O(N \ln(N))$. Readers wanting more details are referred to (Liu et al. 2000).

Multi-level multi-integration Method (MLMI)

Another preferable method for evaluating the deformation is the multi-level multi-integration method. An outline of MLMI algorithm is presented as below.

In (3), the integration kernel function $K(x) = \ln|x|$ is smooth everywhere except the singularity point ($x = 0$). In a numerical analysis, the integration has to be evaluated in discrete form over a grid with the mesh size h

$$v_i^h = h \sum_j K_{i,j}^{hh} p_j^h \quad (10)$$

where v_i^h and p_j^h denote the discrete value of function $v(x)$ and $p(x)$ on the grids with mesh size h , respectively, and the value of $K_{i,j}^{hh}$, which approximates the kernel function $K(x)$, used to be predetermined by analytical approaches (Ai 1993).

If a coarse grid with a mesh size H ($H = 2h$, for example) is employed to evaluate the same integration (3), its discrete form can be written as

$$v_i^H = H \sum_j K_{i,j}^{HH} p_j^H \quad (11)$$

The idea of MLMI method is to calculate the integration v_i^H first on the coarse grid (H), and then to evaluate the v_i^h on the fine grid (h) through an interpolation,

$$v_i^h = I_H^h v_i^H \quad (12)$$

where I_H^h denotes an interpolation operator from the coarse grid (H) to the fine grid (h), whose specific form depends on the choice of the interpolation function. If this process has been applied to a system with several levels of grid, the reduction in computation times can be very significant.

The problem is, however, the evaluation of (12) may create a considerable error due to the singularity of the

kernel function. In these cases, a correction process has to be introduced. Here a two-level grid system with $H = 2h$ is used as an example to illustrate the approach of correction.

1. Calculate the integration (1) on the coarse grid (H) to get v_I^H .
2. For the points $i = 2I$ on the fine grid, the value of v_i^h can be evaluated based on v_I^H , but with the following correction in the neighborhood of $i = j$ if the kernel $K(x)$ supposed to be singular at $x = 0$

$$v_i^h \approx v_I^H + h \sum_{\|i-j\| \leq m} C_{i,j}^{hh} p_j^h \quad \text{if } i = 2I \quad (13)$$

3. Calculate the interpolation $[I_H^h v^H]_i$ for the points where $i = 2I + 1$, and then evaluate v_i^h using the correction

$$v_i^h = [I_H^h v^H]_i + h \sum_{\|i-j\| \leq m} D_{i,j}^{hh} p_j^h \quad \text{if } i = 2I + 1 \quad (14)$$

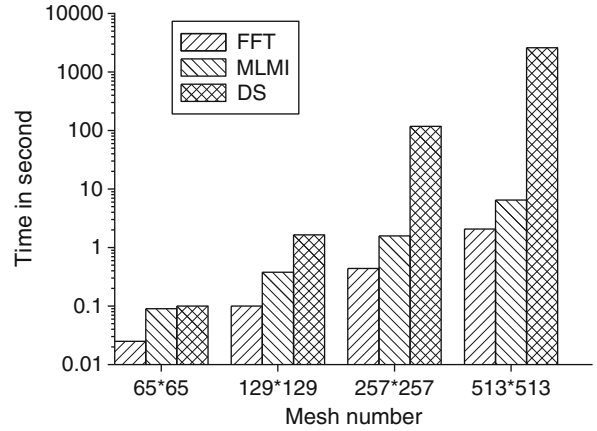
The correction terms $C_{i,j}^{hh}$ and $D_{i,j}^{hh}$ appearing in (13) and (14) have to be precomputed. For more details, refer to (Brandt and Lubrecht 1990).

The above algorithm can be easily extended to two-dimensional cases by applying the same procedure alternately in the x dimension and the y dimension.

Comparison and Discussion

Comparisons of the accuracy and efficiency for three numerical procedures, the direct summation, DC-FFT based method, and MLMI, are made in this section. The three methods were applied to calculate normal surface deformations at different levels of grids, under the load of a uniform pressure on a rectangle area $2a \times 2b$, or a Hertzian pressure on a circle area in radius a . The calculations were performed on the same personal computer, the computational domain was set as $-1.5a \leq x \leq 1.5a$ and $-1.5a \leq y \leq 1.5a$, and covered by a uniform mesh with the node number ranging from 64×64 , 128×128 , and 256×256 to 512×512 for all computational cases.

The CPU times required for the computations on different grids are displayed in Fig. 1. It can be seen that the DC-FFT-based method is the fastest among the three methods. If one multiplication is defined as a unit operation and the total number of nodes is N , the computation in DS method will take N^2 operations, while in theory



Surface Deformation Calculation for EHL, Fig. 1 CPU time for different numerical methods to calculate surface elastic deformation

both computations in MLMI and FFT require $Mn(N)$ operations. In practice, however, the MLMI method performs slower than the FFT, probably due to the complicated programming in MLMI, which causes additional operations.

Concerning the numerical accuracy, the closed-form solutions of normal surface deformation have been compared with the numerical results calculated through the three methods of DS, DC-FFT, and MLMI. The influence coefficients used in the numerical analyses were obtained from bilinear interpolation. The relative errors, as defined in (15), are shown in Fig. 2.

$$\varepsilon = \frac{\sum_i \sum_j |v_{i,j} - v_{i,j}^c|}{\sum_i \sum_j |v_{i,j}^c|} \quad (15)$$

where $v_{i,j}$ and $v_{i,j}^c$ are the deformations obtained through numerical and closed-form solutions, respectively.

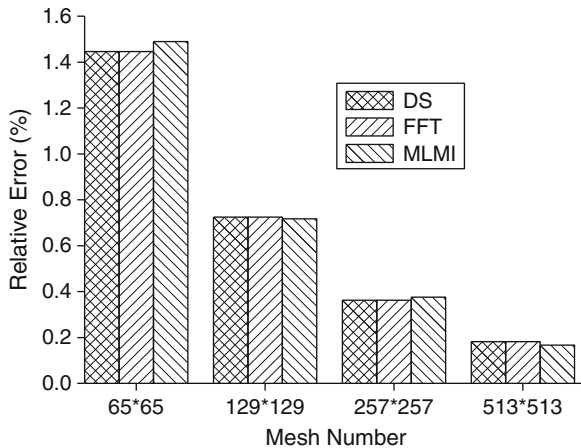
Two conclusions can be drawn from Fig. 2: (a) The numerical accuracy is dominated largely by the discrete errors. When the same influence coefficients were employed, the DS, MLMI, and FFT-based methods produced similar errors, indicating that fast computation methods did not introduce additional errors. (b) The errors decrease roughly by a factor of two if the calculations are performed on a finer grid with half-mesh size.

In summary, both the FFT-based method and MLMI method are expected to be powerful numerical approaches for the deformation evaluation with similar accuracy, especially when a larger number of grid nodes are involved.

Key Applications

Dry Contact Analysis

Dry contact analysis is a very important step to understand the surface interaction, lubrication, and wear. For ideally smooth surface, the contact analysis may need

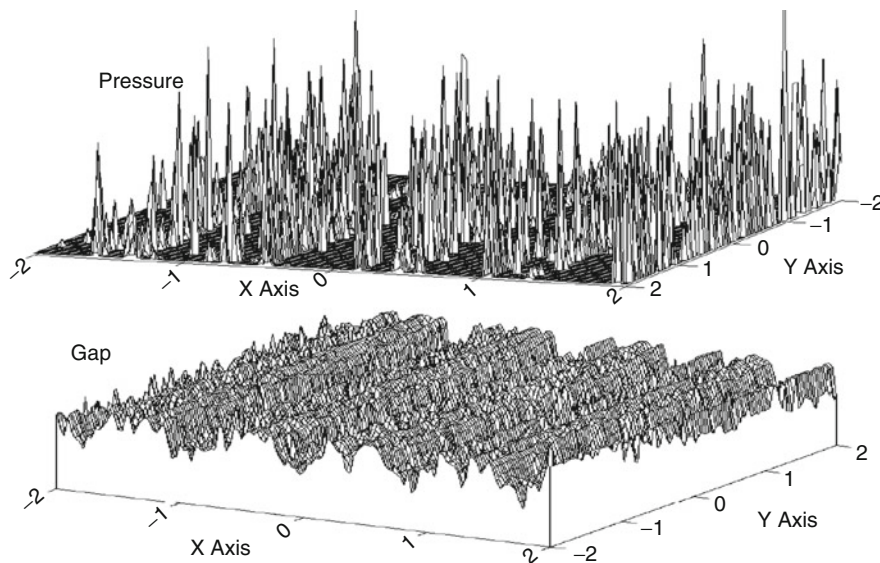


Surface Deformation Calculation for EHL, Fig. 2 Relative errors for a uniform pressure on a rectangle area $2a \times 2b$, in which the multi-summation is calculated via DS, FFT, and MLMI, and IC is determined through a bilinear interpolation-based scheme

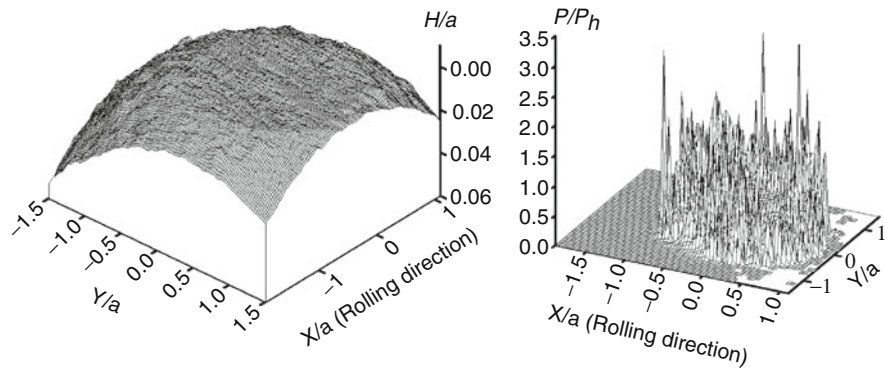
relatively few grid points and thus little computation time. However, the practical engineering surface is rough and the surface topography significantly influences the contact stresses and eventually the operating life of elements. In order to obtain the local contact information, surface topography has to be described with significantly fine resolution, which results in a large amount of grid points; consequently, the numerical simulation requires a large amount of data calculation and considerably longer computation time. The application of fast algorithms of surface deformation will allow a high-speed calculation without sacrificing accuracy. Based on the typical combination of DC-FFT algorithm or MLMI method for surface deformation with a conjugate gradient method, the dry contact analysis can be conducted within several seconds, which enables application in engineering. Figure 3 shows a typical result of dry contact simulation with measured surface based on DC-FFT and CGM.

Lubrication Analysis

The calculation of normal surface deformation plays a key role in numerical solutions of elastohydrodynamic lubricated (EHL) contacts. In the iteration process for pressure distribution in EHL analysis, the bulk of the computation time is spent on the calculation of surface elastic deformation. With the application of fast algorithm of surface deformation, the computation time of numerical analysis for EHL problem can be greatly reduced and satisfy the



Surface Deformation Calculation for EHL, Fig. 3 The results of dry contact analysis based on DC-FFT and CGM



Surface Deformation Calculation for EHL, Fig. 4 Numerical simulation results of an EHL problem under point-contact condition with measured surface

engineering demands. Figure 4 shows the typical numerical simulation results of an point contact EHL problem.

Nomenclature

a, b	Half-length of the contact region in the x and y directions (m)
C_{ij}^{hh}, D_{ij}^{hh}	Correction terms in MLMI
E^*	Effective Young's modulus (Pa)
h	Mesh size of the fine grid
H	Mesh size of the coarse grid
$H(x)$	Response or Green's function
I_H^h	Interpolation operator from coarse grid (H) to fine grid (h)
$K(x)$	Integration kernel function in MLMI
K_j	Influence coefficient at the origin owing to a unit load acting on position x_j
K_j^i	Influence coefficient at point x_i owing to a unit load acting on positing x_j
K_{ij}^{hh}	Coefficient used in MLMI to approximate Green's function on grid h
N, M	Number of the grid in x and y direction
$N(x)$	Interpolation function to approximate the pressure distribution within the element
p	Contact pressure
p_H	Maximum Hertzian pressure (Pa)
v	Normal surface deformation
x, y	Coordinate in the x and y direction
$\Delta x, \Delta y$	Dimensions of an element in the x and y direction
ε	Relative error
$\{\}_n$	Series of size n
Hat (\wedge)	Discrete Fourier transform

Cross-References

- [3D Line Contact EHL](#)
- [Differential Scheme Effect on EHL Solution](#)
- [EHL Governing Equations](#)
- [EHL, Full Numerical Solution Methods](#)
- [Lubricant Non-Newtonian Effect on EHL](#)
- [Mesh Density Effect on EHL Solution](#)
- [Mixed EHL](#)
- [Point Contact EHL](#)
- [Thermal EHL Theory](#)

References

X. Ai, Numerical analyses of elastohydrodynamically lubricated line and point contacts with rough surfaces by using semi-system and multi-grid methods. Ph.D. Thesis, Northwestern University, 1993

A. Brandt, A.A. Lubrecht, Multilevel matrix multiplication and fast solution of integral equations. *J. Comput. Phys.* **90**, 348–370 (1990)

K.L. Johnson, *Contact Mechanics* (Cambridge University Press, Cambridge, 1985)

S. Liu, Q. Wang, G. Liu, A versatile method of discrete convolution and FFT (DC-FFT) for contact analyses. *Wear* **243**, 101–111 (2000)

A.A. Lubrecht, E.A. Ioannides, Fast solution of the dry contact problem and associated surface stress field, using multilevel techniques. *J. Tribol. T ASME* **113**, 128–133 (1991)

C.H. Venner, A.A. Lubrecht, *Multilevel Methods in Lubrication* (Elsevier, Amsterdam, 2000)

W.-Z. Wang, H. Wang, Y.-C. Liu, Y.-Z. Hu, D. Zhu, A comparative study of the methods for calculation of surface elastic deformation. *Proc. Inst. Mech. Eng. J-J Eng.* **217**(J2), 145–153 (2003)

Surface Displacement Calculation

- [Surface Deformation Calculation for EHL](#)

Surface Distress

- [Gear Surface Pitting Failure and Pitting Life Analysis](#)

Surface Durability

- [Gear Surface Pitting Failure and Pitting Life Analysis](#)

Surface Energy

- [Surface Forces, Surface Tension, and Adhesion](#)

Surface Energy and Adhesion

- [Basic Concepts in Adhesion Science](#)

Surface Engineering via Coatings

- [Tribological Coatings for High-Temperature Applications](#)

Surface Ennoblement

- [Tribological Coatings for High-Temperature Applications](#)

Surface Force Apparatus

YU TIAN

State Key Laboratory of Tribology, Tsinghua University,
Beijing, People's Republic of China

Synonyms

[SFA-surface force apparatus](#)

Definition

Surface force apparatuses are research instruments for directly measuring the static and dynamic forces between surfaces, and for studying other interfacial and thin film phenomena at the molecule level.

Scientific Fundamentals

Backgrounds

Several different techniques have been developed for measuring ► [surface forces](#) and intermolecular forces both in air and in liquids: the surface forces apparatus (SFA), the osmotic stress device, the force balance, and the atomic force microscope (AFM). These techniques have allowed accurate measurements of surface forces from the macroscopic to the molecular (subnanoscopic) scale, leading to an improved understanding of these forces as well as their implications in colloidal behavior, adhesion, ► [friction](#), and dynamic (nonequilibrium) interactions (Drummond and Richetti 2007). The SFA was pioneered by Tabor, Winterton, and Israelachvili in the early 1970s at Cambridge University.

The SFA measures the forces between two surfaces in vapors or liquids with a sensitivity of 10 nN and a distance resolution of 0.1 nm. It also can be used to measure the refractive index of the medium between the surfaces, molecular orientations in thin films (under certain conditions), adsorption isotherms, capillary condensation, surface deformations arising from surface forces, dynamic interactions such as viscoelastic and frictional forces, contact and friction of rubbers/elastomers, and other time-dependent phenomena in real time. Though mica surfaces are the primary surfaces used for these measurements, it is possible to deposit or coat these surfaces with surfactants, including lipids, polymers, proteins, metals, metal oxides, and silica, so as to alter the nature and chemistry of the interacting surfaces while keeping them smooth by virtue of the molecularly smooth mica substrate surface underneath.

The surface force apparatus has been developed from SFA Mark I to SFA Mark III; SFA 2000 is easier to operate and is generally more user-friendly. The four distance controls have been specially designed to produce perfectly linear displacements of the surfaces without backlash, and they are more robust, less susceptible to thermal drifts, easier to clean, and require smaller quantities of liquid than old versions. A number of facilities have been built as accessories to allow dynamic measurements of friction, lubrication, and viscoelastic forces over a large range of shear rates and sliding speeds (Israelachvili 1992).

Working Principles

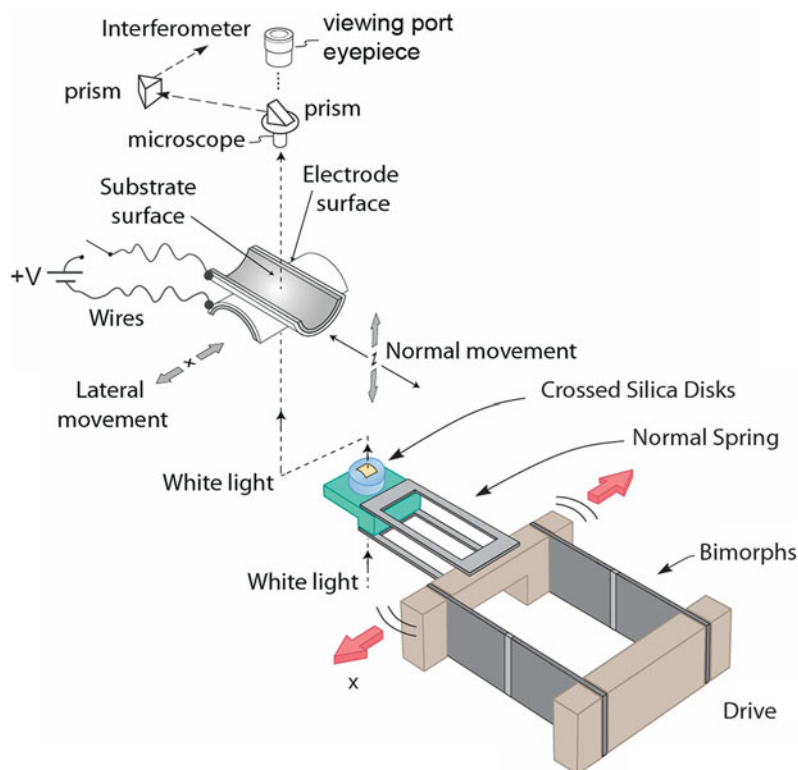
The classical structure of an SFA is shown in Fig. 1. The surfaces tested in SFA are usually two cross-positioned cylindrical mica surfaces. Mica surfaces are molecularly flat and can be covered by other surfactant layers or metallic films. The backside of the mica surface is coated with a semi-reflective layer of silver. The two curved cylinders have the same radius R , and the so-called crossed cylinders geometry is mathematically equivalent to the interaction between a sphere of radius R and a flat surface.

Usually, one surface is fixed on a rigid support, and the other is held by a cantilevered spring that is usually driven to move at a uniform speed. Sometimes, a piezoelectric positioning element is also used to achieve the fine linear motion. Two surfaces are carefully moved towards and retracted from one another, controlled by the rough manual adjustment and the fine motor linear motion. The final several micrometers are usually controlled by fine displacement feeding to contact. The distance between the surfaces can be controlled over a range of 5 mm with a resolution of 0.1 nm by a four-stage mechanism of increasing sensitivity. The stiffness of the force measuring spring can be adjusted during experiments by moving the

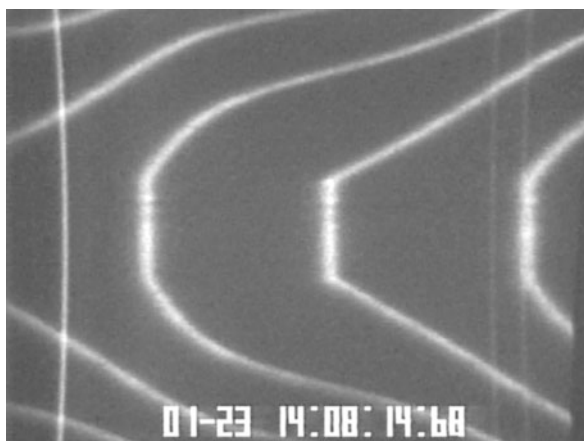
dove-tailed clamp along the length of the spring or changing to a spring with different stiffness during the setup of the experiment.

The shapes and the separation between surfaces are measured with ► [optical interferometry](#) by analyzing the optical interference fringes produced when a parallel white light passes through the two surfaces, called fringes of equal chromatic order (FECO), as shown in Fig. 2. The deflection of the spring is used to calculate the surface force by comparing the driven displacement of the motion and the deflection of the cantilever.

Dynamic measurements are conducted with surfaces in motion by vibrating the piezoelectric crystal supporting the upper surface in the vertical direction. If a friction device of linear motor driven or piezoelectric bimorph is used to laterally shear the surface, a friction test or rheology measurement of thin films can be done. The lower surface is usually fixed on the bimorph; the upper surface is attached to a vertical double cantilever spring, whose deflection is monitored using strain gauges connected to form the arms of a Wheatstone bridge. If a viscous or friction force is acted on the surface by the movement of the lower surface, the strain gauge can sense the deflection



Surface Force Apparatus, Fig. 1 Functional elements of SFA III with an electric field applying attachment



Surface Force Apparatus, Fig. 2 Typical FECO fringes with a flat contact in the middle range

of the spring deformation and the friction force can be tested.

One of the major advantages of SFA is the ability to image the contact area and determine the distance between the surfaces, the refractive index of the film confined between the surfaces, and the geometry of the contact region. It allows one to image in situ and real time geometry of the contact area, and probably is the only technique with sub-nanometer resolution.

Specifications of SFA 2000

SFA2000 has a distance resolution better than 0.1 nm. Positioning accuracy of the four distance controls can reach 200 nm by coarse control (manual differential micrometer), 50 nm by medium control (manual differential micrometer), 0.2 nm by fine control (motorized control of differential spring system), and less than 0.1 nm by the piezo control (Israelachvili 2005).

The sensitivity of measuring absolute force to 1 mdyne between surfaces of radii R is typically between 1 and 50 mm, corresponding to normalized forces (F/R) of 1 $\mu\text{N}/\text{m}$; adhesion, surface, and interfacial energies ($E = F/4\pi R$) of better than 10^{-6} J/m^2 ; and pressures of less than 1 atm.

Maximum pressure: 5,000 atm.

Sliding distance/amplitude during friction experiments (using a friction device and bimorph slider): from several nanometers to 5 mm.

Resolution of measuring lateral, shear, and frictional forces: 1–10 μN .

Resolution of measuring friction coefficients: 0.0001.

Range of shear rates (using a friction device and bimorph slider): 10^{-10} to 10^5 s^{-1} .

Range of lateral speeds (using a friction device and bimorph slider): 10^{-8} to 3 cm/s.

Range of driving frequencies with a bimorph slider: 10^{-6} to 100 Hz.

Developments of SFA

Besides the improvement of Israelachvili, several SFA experimental setups have been proposed during the last two decades. All those proposed are still based on the original method of Israelachvili and Tabor.

The group of Georges et al. developed a SFA in which lateral and normal forces are measured with capacitive force sensors, so they could use non-transparent coating materials. A sphere to a plane geometry is used. The motion of the sphere relative to the plane is also measured with a three-axial capacitance sensor. Steve Granick and coworkers have designed an alternative structure to the measurement of smaller deformations. Jacob Klein and coworkers have also presented a design to greatly improve the sensitivity of friction force measurement. Due to the detection and analysis of FECO fringes, the automation of SFA is difficult. Different strategies of measurement automation have been proposed. Among them, Heuberger and coworkers have done a lot on both the automation and miniaturization of SFA.

In addition to the capability of SFA in determination of the normal or lateral forces between surfaces, the structural information at the molecular level is also desirable in the fundamental research on the surface or interaction where researchers have tried to combine SFA with other techniques in situ to more profoundly disclose the interfacial phenomena. In situ small-angle X-ray scattering has been combined with SFA to give structural information of ultra-thin confined liquid films along with the shear force. Helm and coworkers showed that multi-beam interference (MBI) can be used to obtain structural information of the confined thin films of optically active molecules without any modification of the traditional SFA. The light absorption by the molecules is enhanced by the multiple reflections. Granick and coworkers have coated the back-side surface of mica by multilayer dielectric coatings to obtain different transparency of the optical spectrum as a substitute for the semi-transparent silver layer to apply different spectroscopic tools to test the in situ structural information of thin films.

Other techniques, like fluorescence, confocal Raman spectroscopy, photoluminescence, adsorption dichroism, and quartz crystal resonators can be combined with SFA to provide more information.

Key Applications

Surface Force Measurement

Most universally existing surface forces coming from van der Waals forces can be measured by SFA exposed in air, vacuum, or vapor conditions. Due to the elastic factor of the supporting spring, at a very small distance of several nanometers, the surface fixed on the spring would jump into contact with the fixed surface. Since SFA can image *in situ* and in real time the geometry of the contact area, different contact models can be checked by the apparatus. The relationship among the elastic modulus, the surface energy, and the geometry of spheres can be obtained.

Surface Forces in Vapor or Liquids

Since most biological systems are under wet conditions, the surface forces in solution or in liquid vapor are important to the understanding of biological surface interactions. And in colloidal suspensions like paints and inks, it is important to determine the interaction between particles.

Using the surface forces apparatus, the static and hydrodynamic forces between two substrate-supported lipid bilayers in water and aqueous solutions of poly(ethylene glycol) have been measured. An enhanced adhesion due to depletion attraction was measured at small separations. A repulsive barrier was found at larger separations, which is not predicted by mean-field theories. The discussion on the results in terms of depletion attraction, depletion stabilization, thin film lubrication, the effective viscosity in thin films, and the possibility of polymer aggregates in PEG solutions was carried out. These results are helpful for the understanding of bilayers.

Many biological recognition interactions involve ligands and receptors that are tethered. SFA can be used to directly measure the force-distance interaction between a polymer-tethered ligand and its receptor. At separations near the fully extended tether length, the ligands rapidly lock onto their binding sites, pulling the ligand and receptor together. The measured interaction potential and its dynamics can be modeled with standard theories of polymer and colloidal interactions. This technique can directly measure the interaction force of biological molecules and provide information for the modeling of the related processes.

Friction, Adhesion, and Rheology Test

The friction and adhesion of different surfaces can be readily tested in SFA by modifying the mica surface with the materials to be studied. The friction velocity can be adjusted in a wide range, and the friction force can be

measured with a high resolution, as shown by SFA 2000. Phenomena like static friction force, dynamics friction force, and stick-slip can be observed. The rheology of nanoscale confined fluids is also widely investigated by SFA. Two typical examples are given below. Many details of molecular behavior can be studied by SFA individually, along with other characterizing techniques.

The adhesion force between two molecularly smooth mica surfaces in pure water has been directly measured by SFA as a function of the rotational angle between the two surface lattices. A sharp peak of adhesion has been found due to the variation of the surface energy with the angle rotated between 8° and 180° . Results indicate that the whole interaction potential between the surfaces in liquids may depend on the relative orientation of their surface lattices.

Friction and adhesion hysteresis experiments have been carried out on fluorocarbon surfactant monolayer-coated surfaces using SFA. Measurements were made as a function of temperature, load, sliding velocity, and relaxation time, and the resulting properties are contrasted with those of hydrocarbon monolayers and also with bulk fluorocarbon surfaces. While the overall tribological properties of fluorocarbon surfactants follow the same generic friction phase diagram behavior as hydrocarbon surfactants, the friction phase diagram for fluorocarbon surfactant indicates two different molecular relaxation processes. Chain interdigitation does not play a major role with fluorocarbon surfaces. The surface topography and its change at the molecular and submolecular levels during shear is the most important factor determining the friction of these surfaces.

Cross-References

► [Optical Interferometry](#)

References

- C. Drummond, P. Richetti, Surface forces apparatus in nanotribology, in *Fundamentals of Friction and Wear* (Springer, Berlin/Heidelberg, 2007), p. 15
- J. Israelachvili, *Intermolecular and Surface Forces*, 2nd edn. (Elsevier Academic, Burlington, 1992), p. 169
- J. Israelachvili, *SFA 2000 Manual* (SurForce LLC, USA, 2005), p. 1

Surface Forces

► [Surface Forces, Surface Tension, and Adhesion](#)

Surface Forces in Biosystems

► Adhesion in the Animal World

Surface Forces, Surface Tension, and Adhesion

MARINA RUTHS

Department of Chemistry, University of Massachusetts,
Lowell, MA, USA

Synonyms

Interfacial energy and adhesion; Interfacial forces; Interfacial tension; Surface energy; Surface forces; Surface physics concepts; Surface tension

Definition

Surface forces arise from short-range intermolecular interactions and manifest themselves as longer-range repulsive or attractive, distance-dependent interactions between macroscopic bodies. At a surface, i.e., at a boundary between a condensed phase and a gas, intermolecular attraction causes a net force on molecules away from the surface, toward the bulk condensed phase. As a result, there is an energy “cost” per area, the *surface tension* or *surface energy*, associated with increasing the surface area. When separating two condensed phases along their boundary or interface, new area is created, which requires *work of adhesion* that is related to the magnitudes of the surface energies of the two materials and the interactions across their interface.

Scientific Fundamentals

Surface Forces and Intermolecular Interactions

Interactions between macroscopic bodies across vacuum or a medium arise from the interactions between the constituent molecules of each body across the gap separating them. These intermolecular interactions are the same electromagnetic forces that operate between molecules in a gas, liquid, or solid, determining such bulk properties as non-ideal behavior of gases, boiling and sublimation points, and cohesive strength of solids. Some of the most common intermolecular interactions will be reviewed in the following sections, together with

expressions for the interactions between macroscopic bodies such as colloidal particles, surfaces, and nanoscopic probes of different geometries. van der Waals interactions will be discussed in more detail because of their importance for tribology and adhesion at the nanometer scale (Ruths and Israelachvili 2010).

Interaction forces are obtained from interaction energies by noting that the force is the gradient of the energy, i.e., the derivative of the interaction energy (interaction potential) with respect to separation distance, $F(h) = -dW(h)/dh$.

Van der Waals Interactions

Van der Waals interactions is a collective name for three different, additive interactions responsible for the short-range attraction (cohesion) between atoms and molecules across vacuum or a medium. All three have interaction energies proportional to r^{-6} , where r is the separation distance. van der Waals interactions occur both across vacuum and across a medium.

Van der Waals Interactions Between Molecules in Vacuum (and Inert Vapors)

The dipole–dipole (orientational or Keesom) interaction consists of electrostatic interactions between polar molecules, i.e., permanent dipoles. In liquids and vapors, dipoles are free to rotate but will orient one another so that, on average, an attraction results. The angle-averaged interaction free energy for two freely rotating permanent dipoles in vacuum is

$$w(r) = -\frac{\mu_1^2 \mu_2^2}{3kT(4\pi\epsilon_0)^2 r^6}. \quad (1)$$

where r is the distance between the molecules and μ_i their permanent dipole moments. Equation 1 accounts for the effect of thermal motion, which decreases the mutual alignment of the dipoles (Israelachvili 2011).

The dipole–induced dipole (induction or Debye) interaction arises from dipole moments induced in atoms or in polar or non-polar molecules by the electric field from permanent dipoles, resulting in an attraction between the induced and the permanent dipole. The interaction free energy between two freely rotating permanent dipoles, each with an induced dipole moment α_i , is

$$w(r) = -\frac{\alpha_1 \mu_2^2 + \alpha_2 \mu_1^2}{(4\pi\epsilon_0)^2 r^6}. \quad (2)$$

This interaction does not depend on temperature, since the induced dipole moment will respond to and follow the direction of the permanent dipole moment

when this molecule is reoriented due to thermal motion (Israelachvili 2011).

The induced-dipole–induced-dipole (fluctuation or London dispersion) interaction is present between all atoms and molecules (polar and non-polar) and arises from instantaneous polarization of one atom or molecule due to fluctuations in the charge distribution of a neighboring atom or molecule. This interaction is independent of orientation, but increases with increasing polarizability, giving the initial molecule larger fluctuations in its charge distribution and the responding molecule a larger induced dipole moment. An approximate equation for the interaction energy, assuming only one characteristic vibration frequency of the electrons for each molecule, is

$$w(r) = -\frac{3}{2} \frac{\alpha_1 \alpha_2}{(4\pi\epsilon_0)^2 r^6} \frac{h\nu_1\nu_2}{(\nu_1 + \nu_2)}. \quad (3)$$

where h is Planck's constant and ν_i is the frequency of the lowest electron transition, often approximated by the ionization frequency of an electron in one of the outer orbitals. The largest contribution (70–100%) to the van der Waals interactions comes from the London interaction, except for in the case of highly polar substances, such as water. Since the fluctuations in the charge distributions are rapid, this interaction is weakened by a randomizing effect at large separation distances (>10 nm), a phenomenon called “retardation” (Casimir and Polder 1948; Israelachvili 2011).

At very small separation distances, when the electron clouds of the molecules begin to overlap, all three of these interactions will ultimately be repulsive due to the Pauli exclusion principle. The total van der Waals interaction is thus a combination of very short-range repulsion and a longer-range attraction.

Van der Waals Interactions Between Molecules in a Liquid

Across another material (a medium), the van der Waals forces will be significantly lower than in a vacuum, because the interaction between two solute molecules in a medium (solvent) involves reorientation and displacement of nearest-neighbor solvent molecules, which interact both with the solute and with other solvent molecules. To (partly) account for these solvent effects, ϵ_0 in the equations above is replaced by $\epsilon_0\epsilon_r$, where ϵ_r is the relative permittivity of the (bulk) medium, and α_i is replaced by the excess polarizability of the solute molecule. Although the surrounding medium (solvent) may also change the dipole moment and polarizability of the solute from their values in vacuum, the interaction between two identical molecules remains attractive. The theory for van der Waals

interactions has been expanded to account for more than one absorption frequency in an atom or molecule and the effects of a solvent (Evans and Wennerström 1999; Israelachvili 2011).

Van der Waals Interactions Between Macroscopic Bodies

The contribution of the van der Waals interactions to the interaction energy between two macroscopic bodies (such as colloidal particles, or a nanometer-sized asperity and a flat surface) can be found by summing the pair-wise interactions (1)–(3) of all the molecules in one body with the molecules in the other. The interaction energy was derived by Hamaker for several geometries, including two spheres with identical radii, R , interacting across vacuum (Hamaker 1937; Israelachvili 2011). Similar expressions have been derived for other geometries such as two flat plates and a sphere interacting with a flat plate (or, equivalently, two crossed cylinders), which are surface geometries common in nanotribological measurements:

$$W = -\frac{AR}{12h} \quad (4)$$

two spheres

$$W = -\frac{A}{12\pi h^2} \quad (5)$$

two flat plates

$$W = -\frac{AR}{6h} \quad (6)$$

sphere and flat plate

where A is the Hamaker constant (section “The Hamaker Constant”), and h the closest surface separation ($R \gg h$). The equations above will lead to an overestimation of the dispersion forces at large separations (for bodies interacting across a medium, $h > 10$ nm), since the time needed for propagation of the dipole moments induced by fluctuations is not accounted for (i.e., the retardation effect). By comparing the dependence on separation distance in (1)–(3) to that in (4)–(6) it is, however, apparent that the interaction energy between macroscopic bodies is significantly more long-ranged than that between two molecules.

The Hamaker Constant

The non-retarded Hamaker constant can be calculated from pair-wise interactions (the original “London–Hamaker microscopic approach,” evaluated for one oscillation frequency) or by using the “macroscopic” Lifshitz approach (Lifshitz 1956; Israelachvili 2011), where the interacting

bodies and the intervening medium are treated as continuous phases, and the polarizabilities and dipole moments are determined from bulk dielectric properties. Hamaker constants for solids interacting across vacuum (or air) are around 10^{-20} – 10^{-19} J (the higher values are obtained for metals). When the interacting bodies are separated by a medium (e.g., a liquid) the van der Waals interactions are significantly reduced, and an effective Hamaker constant has to be used. A further approximation (Ninham and Parsegian 1970) of the expression for the Hamaker constant derived from the Lifshitz theory for only one vibration frequency and two different media (1 and 3) interacting across medium 2 is

$$A_{123} = A_{v=0} + A_{v>0} \approx \frac{3kT}{4} \left(\frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1 + \varepsilon_2} \right) \left(\frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_3 + \varepsilon_2} \right) + \frac{3h\nu_e}{8\sqrt{2}} \frac{(n_1^2 - n_2^2)(n_3^2 - n_2^2)}{\sqrt{n_1^2 + n_2^2} \sqrt{n_3^2 + n_2^2} \left\{ \sqrt{n_1^2 + n_2^2} + \sqrt{n_3^2 + n_2^2} \right\}} \quad (7)$$

where the first term ($v = 0$) represents the Keesom and Debye interactions and the second ($v > 0$) the London (dispersion) interaction. ε_i and n_i ($i = 1, 2, 3$) are the static dielectric constants and refractive indexes of the media, and ν_e is the frequency of the lowest electron transition (generally around $3 \times 10^{15} \text{ s}^{-1}$). If the dielectric properties of the intervening medium 2 in (7) are intermediate to those of media 1 and 3, the interactions between 1 and 3 across 2 will be of opposite sign (repulsive, $A < 0$ in (7)) to the interactions between 1 and 1 across 3, and 2 and 2 across 3. For the symmetrical case of two bodies of medium 1 interacting across medium 2 (i.e., $3 \rightarrow 1$ in (7)), A is always positive, i.e., the van der Waals interaction energy between bodies of the same material will be attractive across another material regardless of its dielectric properties (the “like-attracts-like” rule), as is the case for interactions between identical solute molecules in a binary mixture (cf. section “[Van der Waals Interactions Between Molecules in a Liquid](#)”).

Applications of Van der Waals Interactions

Attractive van der Waals interactions are highly important for colloidal (in)stability (cf. section “[Combining Van der Waals and Double Layer Forces: The DLVO Theory](#)”), and for the self-assembly of molecules of technical or biological importance. They also contribute to (and are, in some systems, solely responsible for) the surface energy (tension) of materials (section “[Surface Tension or Surface Energy](#)”) and the adhesion between surfaces (section “[Surface Forces and Intermolecular Interactions](#)”). The existence of repulsive van der Waals interactions between

materials across a medium with intermediate dielectric properties causes a preferential, non-specific adsorption of molecules with an intermediate dielectric constant (e.g., adsorption of medium 2 in the example above to the interface between media 1 and 3). This is commonly observed, for example, as adsorption of vapors or solutes to a solid surface, and contributes, together with intermolecular attraction between adsorbate molecules, to the formation of self-assembled monolayers (SAMs) on solids.

Electrostatic Interactions

Electrostatic interactions (Coulomb interaction) occur between charges that may be present due to ions in a system or because of electron transfer between surfaces in contact (contact electrification or tribocharging). Here, the focus will be on the interaction forces between charged surfaces in contact with a polar liquid, such as an aqueous electrolyte solution. In tribology, such situations are encountered in biological systems and are of importance in systems containing water-based lubricants. Across a liquid, the interaction of charged surfaces is quite complicated, and a number of approximations are typically made.

Electrostatic Double-Layer Forces in a Liquid

When in contact with water or other highly polar liquids, surfaces may become charged, either by dissolution of ions that become solvated and are distributed into the solution due to entropy, or by preferential adsorption of ions from the solution. The charge on the surface is balanced by a decaying concentration of counterions within a thin layer of solution close to the surface. The surface charges and these counterions form a “diffuse double layer.” At low ionic strength (i.e., in dilute solution) a measure of the thickness of this layer is the Debye length (Debye screening length), κ^{-1} . An analogy can be made between the diffuse double layer and a charged capacitor, where κ^{-1} is the distance between plates with the same charge density (one positively and one negatively charged) as the surface in the double layer. The Debye length is thus a measure of the thickness of the “atmosphere of counterions” near the surface (in excess of the concentration in bulk solution), and depends on the ionic strength of the solution.

$$\kappa^{-1} = \sqrt{\varepsilon_0 \varepsilon_r k_B T / \left(e^2 N_A \sum_i z_i^2 c_i \right)} \quad (8)$$

where e is the electron charge, N_A is Avogadro’s constant, z_i is ionic charge, and c_i solution concentration (Evans and Wennerström 1999; Israelachvili 2011). The Debye length

is a property of the solution, not of the charged surface, and is large in dilute solution (for example, 30 nm in 10^{-4} M 1:1 electrolyte and about 1 μm in pure water).

The thickness of the diffuse double layer determines the range of the interaction force between the two surfaces. As the surface separation is decreased, the counterions stay between the surfaces because of their attraction to the surface charges, but their concentration is increased, and repulsion arises due to the increased osmotic pressure in the confined thin film. The so-called electrostatic “double layer force” is thus of entropic origin. At large separation distances h (weak overlap between the ion number densities and thus the potentials), the interaction energy between two similarly charged molecules for surfaces is typically repulsive and decays exponentially with h , the decay length being the Debye length, κ^{-1} .

$$W(h) = \text{constant} \times e^{-\kappa h} \quad (9)$$

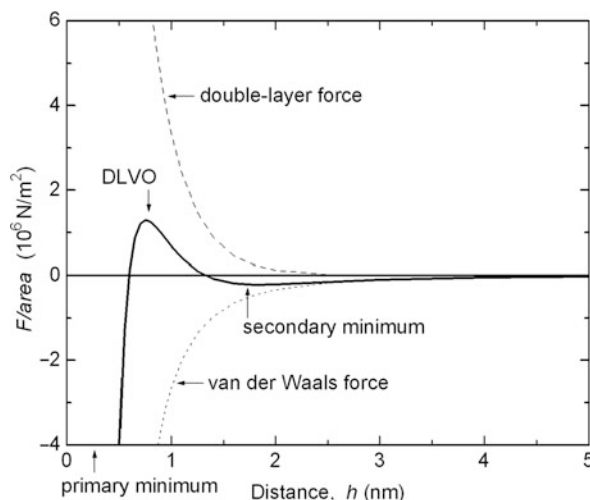
where the constant depends on the geometry of the interacting surfaces, on the surface charge density, and the solution conditions (temperature and ionic strength). The constant is found by solving the Poisson–Boltzmann equation under different constraints and approximations described in detail in textbooks on colloid science, where solutions for different geometries are provided (Evans and Wennerström 1999; Israelachvili 2011).

The equations above are accurate only at large separation distances, i.e., at separations larger than about one Debye length. At very small separations, it can be shown that the interaction potential energy depends on whether the surfaces remain at constant potential (as assumed above), or at constant charge density (which would result in a larger repulsion than that in the equations above), or somewhere between these limits. At “constant charge,” the total number of counterions in the compressed solution (gap) remains constant, whereas at constant potential, the concentration remains constant. The limiting force per area (the so-called “disjoining pressure”) at constant charge is simply the osmotic pressure of the confined ions.

There are cases where the double-layer force can be attractive at small separation distances although the surfaces are of similar charge. If the surface contains chargeable chemical groups, these can dissociate (charge regulation), or ions can condense from solution onto the surfaces to lower the effective surface charge density. Especially at high charge density, the mobility and polarizability of the many counterions close to the surface can decrease the repulsion (ion correlation or charge fluctuation) (Evans and Wennerström 1999; Israelachvili 2011).

Combining Van der Waals and Double Layer Forces: The DLVO Theory

At small separation distances, all surfaces experience van der Waals interactions. Between solid surfaces immersed in a polar solvent, these are quite strongly reduced, but become important at small separations where they, because of their different distance-dependence, can overcome the repulsive double-layer forces. This is the theoretical prediction behind the so-called DLVO theory (Derjaguin–Landau–Verwey–Overbeek) (Verwey and Overbeek 1948; Evans and Wennerström 1999; Israelachvili 2011). If the van der Waals interactions and the double-layer force are assumed to be additive, the total force–distance curve can show several maxima and minima, where the outer or secondary minimum (Fig. 1) typically is not more than a few $k_B T$ deep and is responsible for the (reversible) flocculation of colloidal particles from an aqueous suspension. Particles interacting at a separation distance corresponding to the inner or primary minimum experience irreversible coagulation. For surfaces of a given material, the van der Waals force is not easily altered, but the repulsive maximum in the force between these two minima can be lowered by lowering the repulsive double-layer force, i.e., by altering the charge density of the surfaces or, what is often easier, by increasing the ionic strength of the solution by choosing different ions (different valency) or changing the concentration.



Surface Forces, Surface Tension, and Adhesion, Fig. 1 DLVO interaction (solid curve) between two flat surfaces, calculated as the sum of the attractive van der Waals force (dotted curve, $A = 5 \times 10^{-20}$ J) and repulsive electrostatic double layer force (dashed curve, surface potential 100 mV, 1 M 1:1 electrolyte, $\kappa^{-1} = 0.3$ nm)

Applications of Electrostatic Double-Layer Forces

Double-layer forces are commonly present between surfaces in aqueous solution and can be long-ranged. They are of importance for colloidal stability (for example, for the stabilization of suspensions and association colloids) and are encountered in many biological systems.

In practice, there are also other forces than van der Waals and double-layer forces at small separation distances, and a very highly confined thin film can have more features that cause additional repulsive barriers than the one in the DLVO theory. In addition to electrostatic double-layer (i.e., entropic) and van der Waals forces, the interactions between particles are influenced by structural or solvation forces, arising from positional ordering (layering) of spherical or unbranched molecules or aggregates between smooth surfaces. Solvation forces are oscillatory functions of surface separation, with oscillation periods corresponding to the thicknesses of layers of semi-ordered molecules or aggregates. The adsorption of large molecules on a surface can also strongly affect the interaction forces, as discussed below.

Polymer-Mediated Interactions in Solution

If molecules protrude from surfaces, they may cause a steric repulsion due to compression and osmotic effects as the surfaces come close together. This is common in systems containing adsorbed polymers. Polymers or macromolecules are long chains of segments (monomers) linked together by covalent bonds. In solution, they form coils whose size depends on the strength of interaction between the segments and the solvent, the so-called solvent quality, where a “good solvent” is a situation where the interactions (van der Waals and hydrogen bonding) between segments and solvent is stronger than the segment–segment attraction. This causes an expansion of the polymer coil in solution and an extension into solution of chains adsorbed to a surface. The opposite case is a poor solvent. A solvent with the same strength of interaction with the segments as the segment–segment attraction is a theta solvent for the polymer. If the adsorbed amount of polymer is low, the attraction may become attractive due to bridging, i.e., adsorption of a polymer chain to two opposing surfaces. In the case of non-adsorbing polymer, an attractive force can arise as a result of the osmotic pressure of the polymer in solution (section “[Depletion Interactions](#)”).

Interactions Between Adsorbed Homopolymer Layers

Homopolymers (polymers containing only one type of segment) interact with surfaces through van der Waals

and electrostatic interactions. The physisorption of homopolymers is reversible and highly dynamic, i.e., the adsorbed and free segments exchange rapidly, but the exchange with free chains in the solution is slow, since the polymer remains bound to the surface as long as one segment along the chain is adsorbed. The adsorption energy per segment is generally on the order of $k_B T$, and a polymer chain is said to adsorb as trains (segments in contact with the surface), loops, and tails (freely dangling chain ends). Scaling theories for semidilute good solvent conditions predict that the adsorbed layer is proportional to the extension of the longest loops, i.e., to the polymer coil size in solution, which is proportional to $M_w^{0.6}$, where M_w is the weight average molecular weight.

Interaction forces between adsorbed polymer layers arise from a balance between intermolecular forces and the entropy of mixing. Theoretically and experimentally, there are two distinctly different types of equilibrium interactions between adsorbed polymer layers: “true” equilibrium, where polymer is desorbing and migrating (diffusing) out from the confining gap between the surfaces into the surrounding bulk solution, where it can gain conformational entropy, and “restricted” equilibrium, where the total amount of confined polymer is kept constant (unchanged from the equilibrium amount adsorbed on the surface before the confinement), but the distribution of adsorbed and non-adsorbed polymer segments within the gap is changing as the surfaces are brought together.

Experimentally, it is difficult to reach a true equilibrium situation, since the time needed for total desorption of a high molecular weight polymer is very long, especially when confined between two surfaces. In general, some kind of restricted equilibrium is investigated. Theoretically, at restricted equilibrium, the total amount of polymer between the surfaces equilibrates across the gap, i.e., the individual polymer chains are no longer only associated with the surface they were originally adsorbed on. Experimentally, even this condition is difficult to reach, and one often measures the interactions between two adsorbed, separate polymer layers, which is an even more restricted condition than that assumed in the theory for restricted equilibrium. At “true” equilibrium, some theories for interactions between adsorbed layers of randomly adsorbed homopolymer predict a monotonic attraction due to bridging and depletion interactions (cf. section “[Depletion Interactions](#)”) as the surfaces are brought closer (de Gennes 1982; Scheutjens and Fleer 1985). Bridging occurs when segments from a chain adsorbed on one surface are able to reach over to another surface and adsorb on it. The polymer chain would then gain conformational entropy if the two surfaces came

closer together, and the result is a net attractive force acting between the two surfaces. In addition, in a poor solvent, there are attractive segment–segment interactions.

At restricted interaction equilibrium, scaling theories for interactions between adsorbed layers under good solvent conditions predict a monotonic (steric-entropic) repulsion (de Gennes 1982). This is commonly observed for high molecular weight polymers, which are used for steric stabilization of colloidal suspensions. In mean-field theories, the configuration-dependent interaction potential in scaling theories is replaced with a mean potential resulting from the distribution of chain configurations (Scheutjens and Fleer 1985). The predicted interaction varies from attractive (bridging) at low adsorption density (produced from low segment concentration in the solution) to repulsive at high concentration. The range of the interaction is predicted to be dominated by the extension of loops.

Polymer Brush Interactions

If an end group or a “block” of segments within the polymer chain is different from the rest of the chain, it may preferentially adsorb on the surface due to electrostatic interaction, chemical bonding, or different solubility in the solvent. End-adsorbed polymers are attached to the surface in only one point, and the extension of the chain, which arises from a balance between elastic energy and conformational entropy, is dependent on the grafting density, i.e., the average distance s between adsorbed end-groups on the surface. At low coverage, where there is no overlap between neighboring chains, the thickness of the adsorbed layer will be proportional to $M_w^{0.5}$ in a theta solvent and to $M_w^{0.6}$ in a good solvent. At higher grafting densities, the chains avoid overlapping one another and will thus extend further into solution so that the adsorbed layer thickness is proportional to M_w . One of the earliest models for the interactions of monodisperse brush systems (Alexander 1977; de Gennes 1987) assumes that the polymer density profile is a step function. In this approximation, the force between a sphere and a flat surface covered with polymer brush is

$$\frac{F(h)}{R} = \frac{16\pi k_B T L_0}{35 s^3} \left[7 \left(\frac{2L_0}{h} \right)^{5/4} + 5 \left(\frac{h}{2L_0} \right)^{7/4} - 12 \right] \quad (10)$$

where L_0 is the height of the brush. The first term arises from the osmotic pressure and the second from chain elasticity. Mean field theories developed later have shown that the density profile is parabolic (Milner et al. 1988;

Zhulina et al. 1990), and a model for polydisperse brushes has also been developed (Milner et al. 1989).

The main differences in the interactions between polymer brushes and adsorbed homopolymer layers is the reduction or absence of hysteresis in the force–distance curve on approach and separation of brush layers, and the lack of interpenetration and entanglement due to the strong stretching of the chains. Bridging is also avoided with polymer brush layers, which are therefore of practical importance for modification of surface properties. Modeling of the conformation of polymers at surfaces and the interactions of adsorbed homopolymer layers and brush layers has attracted significant interest.

Depletion Interactions

In systems containing sufficiently high concentrations of nonadsorbing polymers or large aggregates of self-assembled molecules (micelles), an attractive force may arise at small surface separations (smaller than the coil or aggregate diameter) from the difference in osmotic pressure of the solution remaining in the gap between the surfaces and the surrounding (bulk) solution containing a high concentration of polymer. This phenomenon is called depletion attraction and is proportional to the number concentration of the solute. In highly concentrated systems, one may also observe depletion stabilization, i.e., a weakly repulsive force at a separation distance of about twice the polymer coil diameter. Albeit found over a different distance regime, depletion interactions are similar to the solvation (oscillatory) forces mentioned at the end of section “Applications of Electrostatic Double-Layer Forces”, where alternating repulsive barriers and attractive minima can be observed as the separation distance is decreased such that only discrete numbers of semi-ordered solvent layers can be accommodated between the surfaces.

Applications of Polymer-Mediated Forces

Polymer layers on surfaces are often used to stabilize suspensions of colloidal particles against the coagulation due to van der Waals forces seen in Fig. 1. If bridging can be avoided, as is typically the case with dense polymer layers (especially polymer brushes), a steric repulsion arises due to compression and osmotic effects as the particles come close together. The dielectric properties of the polymer may also be chosen to be similar to those of the solvent, which strongly reduces the van der Waals attraction (“refractive index matching”), even to a point where the particles stay dispersed due to thermal motion. In tribological applications, polymer brush layers are of importance for the prevention of adhesion and wear, and

layers of charged polymers, polyelectrolytes, have recently been found to give ultralow friction of potential importance for the lubrication of biological systems (e.g., joints) (Raviv et al. 2003).

The Derjaguin Approximation

The Derjaguin approximation (Derjaguin 1934) relates the interaction energy between two flat plates to the interaction force between two spheres or between a sphere and a flat surface (which is, from an experimental point of view, more easily achievable than measurements of interactions between flat plates). For a sphere near a flat surface, or two crossed cylinders of radius R , the force is given by

$$F(h)_{\text{curved}} = 2\pi RW(h)_{\text{flat}} \quad (11)$$

This relation was originally derived for an inverse power-law pair-potential, but has been shown to be valid for any type of interaction. Note that the interaction force between two flat plates is found by taking the derivative of $W(h)_{\text{flat}}$ with respect to the separation distance, h , and the interaction energy between a sphere and a flat surfaces is obtained by integrating over $F(h)_{\text{curved}}$.

Surface Tension and Wetting

Surface Tension or Surface Energy

A surface is commonly defined as the region between a condensed phase and a gas or vapor, and an interface as the region between two condensed phases, although these terms are sometimes used interchangeably. Surface tension or surface (free) energy arises from the short-range forces that are responsible for the liquid state of materials, i.e., van der Waals forces (in particular, dispersion forces that are always present), hydrogen bonding, and metal bonding. Molecules in a bulk liquid experience these forces in all directions, whereas the ones located at a surface experience an unbalanced attraction toward the condensed phase. The result is a minimization of the surface area, which is the reason for the spherical shape of small gas bubbles and droplets of liquids. The surface tension, γ , is the work required to reversibly create a unit area at constant temperature, in units of mJ/m^2 , mN/m , or dyn/cm (the surface tension of liquids is often thought of as a force acting perpendicular to a line of unit length on the liquid surface). Strong intermolecular forces lead to high surface tensions or energies, as shown in Table 1.

At an interface, an imbalance of the intermolecular forces occurs as well, but typically with a smaller magnitude than at the condensed phase–vapor surface. Typically, the interfacial tension or interfacial (free) energy lies

Surface Forces, Surface Tension, and Adhesion, Table 1
Surface energies (surface tensions) of different liquids at 20°C (Haynes 2011)

Liquid	γ (mJ/m^2)	Intermolecular forces
<i>n</i> -pentane	18	Dispersion interaction (only)
Water	73	Dispersion, dipole–dipole, and dipole–induced-dipole interactions, and hydrogen bonding
Mercury	476	Dispersion interaction and metal bonding

between the surface tensions of the individual condensed phases. Surface and interfacial tensions (energies) are given as positive values irrespective of the chosen sign conventions of the work of adhesion (cohesion) or the interaction energies (where attraction often is given as a negative value). Surface and interfacial energies typically decrease with increasing temperature as a result of decreasing strength of the intermolecular interactions.

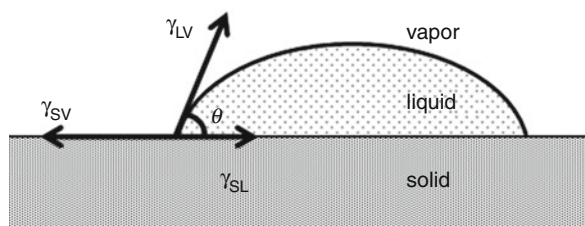
Phenomena arising from surface energies include capillary rise and meniscus formation of condensing liquids (and the related capillary forces that may cause strong attraction between objects from the nanometer to macroscopic scale), nonspecific adsorption, drop shapes, and surface properties such as “hydrophilicity” and “hydrophobicity.” Most of these phenomena are related to the wetting and spreading of liquids on solid or liquid surfaces, as described below.

Spreading, Wetting, and Contact Angle

A commonly observed example of the effects of surface energy is that of a droplet of a liquid (for example, water) interacting with a solid. The liquid will spontaneously spread on the solid if this causes a reduction of the energy of the system, i.e., if the surface energy of the newly formed areas at the bottom and top of the liquid film, $\gamma_{\text{SL}} + \gamma_{\text{LV}}$, is lower than that of the initial solid–vapor surface γ_{SV} . The spreading parameter S is defined as

$$S = \gamma_{\text{SV}} - (\gamma_{\text{SL}} + \gamma_{\text{LV}}) \quad (12)$$

where if $S \geq 0$, the liquid will spread over the solid surface. The case of $S < 0$ is called partial wetting, where the liquid forms a drop with a finite angle with the solid surface called the contact angle (θ in Fig. 2). Similar arguments can be formulated for two immiscible liquids in contact with a solid, or three immiscible liquids. For a liquid in



Surface Forces, Surface Tension, and Adhesion, Fig. 2

Contact angle of a liquid on a solid surface

contact with a solid surface in vapor, the contact angle θ is described by Young's equation

$$\gamma_{SV} = \gamma_{SL} + \gamma_{LV} \cos \theta \quad (13)$$

which can be derived either by minimizing the energy of the system when a drop of liquid is interacting with a solid across air and adsorbing on the surface (Israelachvili 2011) or by balancing the components of the three surface energies at the three-phase line in the direction parallel to the surface (cf. Fig. 2) (Evans and Wennerström 1999; Israelachvili 2011). In practice, γ_{SL} is seldom known, which limits the direct applications of Young's equation. Various approaches have been developed to extract surface energies of solids from measured contact angles by making certain approximations and by comparing results obtained with liquids of different polarity (see section "Work of Adhesion and Surface Energy").

The wetting of a (solid) surface is strongly influenced not only by the surface properties but also by small-scale surface roughness and chemical heterogeneity, which can lead to enhanced contact angle hysteresis, i.e., an enhancement of the difference between the contact angles measured for a drop advancing over the surface and the same drop retracting as the liquid is withdrawn. Some contact angle hysteresis is also observed on smooth surfaces for chemically homogeneous systems and is related to hysteresis in adhesion energy.

Applications of Surface Tension/Energy and Wetting

Surface energy and wetting phenomena are ubiquitous in tribology, especially in applications at the nanometer and micrometer scale. Between surfaces that are wetted or partially wetted by liquids, capillary bridges may form from the vapor phase and give rise to unwanted, strongly attractive forces, for example, in lubrication of microelectromechanical systems and computer hard drives. Typically, these capillary forces are stronger than the van der Waals forces at distances below a few nanometers.

Various surface coatings are applied to minimize this effect by altering the surface energy of the solid surface, and patterning techniques are employed to reduce the bridge formation. Even in the absence of a capillary bridge, a high surface energy of interacting materials leads to a high adhesion between them across a material with lower surface energy.

Work of Adhesion and Surface Energy

Adhesion between surfaces can arise from several different phenomena, such as interlocking of features on the surfaces or interdiffusion of polymer chains, formation of chemical bonds, tribocharging (which gives rise to a strong electrostatic attraction), and van der Waals interactions, which may be negligible in systems where any of the abovementioned phenomena occur, but can be solely responsible for the adhesion of materials where no other forces are present. This section focuses on adhesion arising from surface energy due to van der Waals forces and hydrogen bonding.

When separating two blocks of the same material (material 1) over a unit area, two new unit areas are created. The energy associated with this process, the work of cohesion, is defined as

$$W = 2 \gamma_1, \quad (14)$$

whereas for the separation of different materials (materials 1 and 2), the work of adhesion is the energy associated with creating one new unit area surface of each material, minus the energy associated with the 12 interface, as given by the Dupré equation

$$W_{SL} = \gamma_1 + \gamma_2 - \gamma_{12} \quad (15)$$

Assuming that material 1 is a solid and material 2 a liquid, and combining (15) with (13), one obtains the Young–Dupré equation (Evans and Wennerström 1999; Israelachvili 2011)

$$W_{SL} = \gamma_{LV}(1 + \cos \theta) \quad (16)$$

A large number of approximations have been developed in order to extract surface energy values from contact angle measurements using the Young–Dupré equation (Erbil 1997). One example is the approach by Fowkes, where it is assumed that the work of adhesion arises only from dispersion forces, $W_{SL} = \sqrt{\gamma_S^d \gamma_L^d}$, where γ_S^d and γ_L^d are the dispersive components of the surface energies of the solid and liquid, respectively. This has been shown to hold for the interactions of a nonpolar solid (for example, a hydrocarbon surface) with water, and shows that the interaction of water with hydrocarbon is dispersive.

For polar solid surfaces, more complex approximations have been made, for example, the Owens–Wendt approximation

$$W_{SL} = \sqrt{\gamma_S^d \gamma_L^d} + \sqrt{\gamma_S^p \gamma_L^p} \quad (17)$$

where γ_S^p and γ_L^p are the polar components of the surface energies of the solid and liquid, respectively (and the total surface energies are $\gamma_S = \gamma_S^d + \gamma_S^p$ and $\gamma_L = \gamma_L^d + \gamma_L^p$). The components γ_S^d and γ_S^p can be extracted from contact angle data obtained using several liquids with different polarity (i.e., different values of γ_L^d and γ_L^p). A large body of literature is available on various approximations taking into account hydrogen bonding and acid–base properties (Erbil 1997). Because of the relative ease by which contact angles can be measured, this is a common technique to extract information about the surface energy of surfaces.

The total force between two surfaces in contact, and thus the adhesion between them, is strongly dependent on the structure (roughness) of the surfaces at the nanoscopic level (Ruths and Israelachvili 2010), i.e., in the separation range where the strongest effects are expected from many of the surface forces above.

Cross-References

- Basic Concepts in Adhesion Science
- Capillary Force and Surface Wettability
- Interfacial Energy
- Liquid Contact Angle Measurement
- Polymer Adhesion
- Surface Force Apparatus
- Surface Free Energy
- Van der Waals forces
- Work of Adhesion and Work of Cohesion

References

- S. Alexander, *J. Phys. (Paris)* **38**, 983–987 (1977)
- H.B.G. Casimir, D. Polder, *Phys. Rev.* **73**, 360–372 (1948)
- P.G. de Gennes, *Macromolecules* **15**, 492–500 (1982)
- P.G. de Gennes, *Adv. Colloid Interface Sci.* **27**, 189–209 (1987)
- B.V. Deryagin (Derjaguin), *Kolloid-Z.* **69**, 155–164 (1934)
- H.Y. Erbil, Surface tension of polymers, chapter 9, in *Handbook of Surface and Colloid Chemistry*, ed. by K.S. Birdi (CRC Press, Boca Raton, 1997), pp. 265–312
- D.F. Evans, H. Wennerström, *The Colloidal Domain – Where Physics, Chemistry, Biology and Technology Meet*, 2nd edn. (Wiley-VCH, New York, 1999)
- H.C. Hamaker, *Physica* **4**, 1058–1072 (1937)
- W.M. Haynes, Ed. *CRC Handbook of Chemistry and Physics*, 91st edn. (CRC Press/Taylor and Francis, Boca Raton, 2011).
- J.N. Israelachvili, *Intermolecular and Surface Forces*, 3rd edn. (Elsevier, Amsterdam, 2011)
- E.M. Lifshitz, *Sov. Phys. JETP (English Translation)* **2**, 73–83 (1956)
- S.T. Milner, T.A. Witten, M.E. Cates, *Macromolecules* **21**, 2610–2619 (1988)
- S.T. Milner, T.A. Witten, M.E. Cates, *Macromolecules* **22**, 853–861 (1989)
- B.W. Ninham, V.A. Parsegian, *Biophys. J.* **10**, 646–663 (1970)
- U. Raviv, S. Giasson, N. Kampf, J.-F. Gohy, R. Jérôme, J. Klein, *Nature* **425**, 163–165 (2003)
- M. Ruths, J.N. Israelachvili, Surface forces and nanorheology of molecularly thin films, chapter 29, in *Springer Handbook of Nanotechnology*, ed. by B. Bhushan, 3rd edn. (Springer, Berlin/Heidelberg, 2010), pp. 857–922
- J.M.H.M. Scheutjens, G.J. Fleer, *Macromolecules* **18**, 1882–1900 (1985)
- E.J.W. Verwey, J.T.G. Overbeek, *Theory of the Stability of Lyophobic Colloids*, 1st edn. (Elsevier, Amsterdam, 1948)
- E.B. Zhulina, O.V. Borisov, V.A. Priamitsyn, *J. Colloid Interface Sci.* **137**, 495–511 (1990)

Surface Free Energy

JUNYAN ZHANG

State Key Laboratory of Solid Lubrication, Lanzhou
Institute of Chemical Physics, Chinese Academy of
Sciences, Lanzhou, People's Republic of China

Synonyms

Contact angle; *CTCP*—constant temperature and constant pressure; *SFE*—surface free energy; *SSFE*—specific surface free energy; Surface tension and surface free energy; Wetting ability; Young's equation

Definition

Surface free energy quantifies the amount of work to create a surface. A fundamental characteristic of liquid surfaces is shrinkage to form rough liquid balls, such as mercury pearls, a bead on lotus leaf, or the autoshrink of a liquid sheet, which are led by surface tension and surface energy (Adam 1941; Calvert 2007; de Gennes 1985; de Gennes et al. 2002). For a liquid, the surface tension (force per unit length) and the surface energy density are identical.

Scientific Fundamentals

Surface Tension

Surface tension is a term quantifying the energy required to disrupt intermolecular bonds when a surface is created (Adam 1941). In the physics of solids, surfaces must be intrinsically less energetically favorable than the bulk of a material; otherwise there would be a driving force for surfaces to be created, and surface is all there would be (Adam 1941). The surface energy may therefore be defined as the excess energy at the surface of a material compared with the bulk.

An example explains the concept: a metal wire or glass rod is employed to bend into the pane, with one of sides able to move freely, marked as CD, and the length of moveable wire is presumed as l , shown as Fig. 1. As a liquid film is placed on the wire frame, the system can reach balance only when a proper drawing force (F) is applied to the moveable wire because the liquid film tends to push the movable wire away. This drawing force is the surface tension of the liquid film (Adam 1941). For a certain liquid at a given physical condition, the value of F is directly proportional to the wire length of l , $F = 2 \times l \times \gamma$, γ is the force that is along with liquid surface and makes the surface shrink in a unit of length, called *coefficient of surface tension* or *surface tension*. The commonly used units are mN/m and dyn/cm, $1 \text{ dyn/cm} = 1 \text{ mN/m}$.

Surface Free Energy

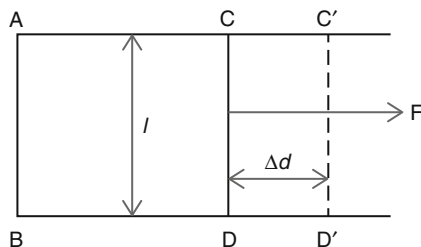
With the system described in Fig. 1, the liquid film tends to expand the area under extra force, meanwhile, the force does work to the system. In a reversible process, while $F = 2l\gamma$, it consumes the least work. At constant temperature and constant pressure (CTCP), the work equals the increase in Gibbs free energy $\Delta G = \gamma \times 2l \times \Delta d$. $2l \times \Delta d$ is the value of the changing in surface area, so (Adam 1941)

$$\gamma = \frac{\Delta G}{2l \cdot \Delta d}$$

γ is the increment of Gibbs free energy with the increasing of surface area per unit square at CTCP, called *specific surface free energy* (SSFE), and it is also called *surface free energy* (SFE). The commonly international used units are J/m^2 , mJ/m^2 , erg/cm^2 , cal/cm^2 , and cal/m^2 . $1 \text{ erg/cm}^2 = 10^{-3} \text{ J/m}^2 = 1 \text{ mJ/m}^2 = 2.39 \times 10^{-8} \text{ cal/m}^2$.

Methods to Measure the Surface Energy

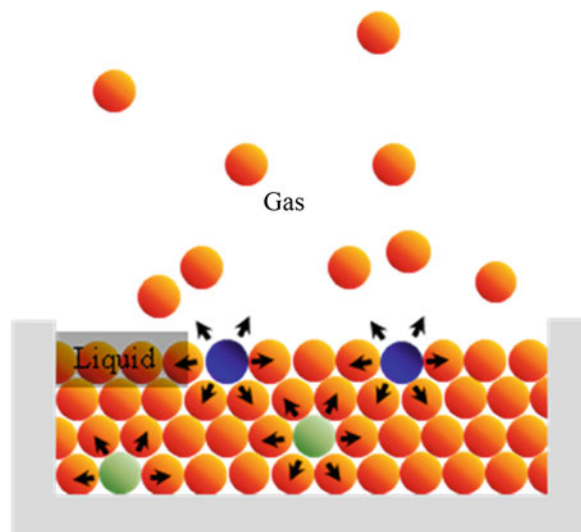
In the early nineteenth century, Laplace pointed out that the surface tension depends on two facts: (1) molecules'



Surface Free Energy, Fig. 1 Model of the surface tension of liquid film

interaction takes place at a distance and the density of the gas phase is much smaller than that of the solid phase. For a century, theoretical research on surface tension and surface free energy has progressed rapidly, but the two points mentioned above retain their fundamental meaning. Figure 2 displays the different intermolecular force of the bulk and the surface of liquid. For the bulk molecules, because intermolecular force generates within a certain distance and the force of all directions are equivalent, the total acting forces are totally offset, that is to say, the total acting force equals zero (0 N). As a result, the bulk molecules can move freely. On the other hand, because the density of the gas phase is much less than that of the liquid phase, the acting force of face molecules is not zero, resulting in inner-directed dragging force. As a consequence, the migration of bulk molecules to the surface must cost energy. Obviously, in a given system, the more face molecules there are, the higher the surface energy is.

The production of surface energy and surface tension differs with variation in matter components and physical conditions (Emil 2008). Among the interatomic and molecular forces, chemical and metallic bonds are stronger and strain the relative displacement of atoms and molecules to form solids (Förch et al. 2009). These bonds always contribute to higher solid surface energy and the value is usually between several hundreds and more than 1,000 mN/m. For common liquids, the interaction between molecules are mainly physical, namely van



Surface Free Energy, Fig. 2 Molecular strained condition at inner and surface of liquid

der Waals force (Adam 1941); the minority have a metallic bond, like mercury. In other liquids, such as water and alcohol-associated solutions, hydrogen bonds mainly contribute to the surface tension and surface free energy, causing the associated solution surfaces to have higher surface energy than the common ones (Moore 1962).

Superposition of Pair Potential

Fowler and Guggenheim, Hamaker, Fowkes, Young, and Crowell et al. have developed a theory of energy increase due to the superimposed molecular potential increment per unit. In 1968, Padday and Uffindell improved the method and calculated a series of normal paraffin isomerization that fitted well with the values of the experiments.

The method assumed that the process of new surface formation expands the two molecule layers' distance to a finite distance. Thus, the fact that work is done against the attractive forces between molecules is ensured and the increased value of energy corresponds to the surface energy of the new surface. This value equals the total energy of the attractions between molecules of the two parts. It can be calculated by adding the potential energy of all molecule pairs, namely superposition of pair potential. The specific approach is discussed below.

Assume there are N junior units (molecule or radical) in one unit volume. Two units' attraction satisfies the van der Waals relationship. If the two units' distance is $r + x$, the interaction energy is described as (White 1948)

$$U_2 = -A/(r + x)^6$$

A is a constant of van der Waals force. There is a shell between $(r + x)$ and $(r + x) + d_x$ to the liquid surface; the attractive energy of the shell to the up-top unit is described as

$$U_2 = -2\pi(r + x)^2 \left(1 - \frac{r}{(r + x)}\right) \frac{r}{(r + x)^6} N d_x$$

The attraction energy of the whole underpart of the liquid to the up-top unit is described as

$$U_2 = - \int_0^{\infty} 2\pi N (r + x)^2 \left(1 - \frac{r}{(r + x)}\right) \frac{r}{(r + x)^6} d_x = \frac{\pi A N}{6r^3}$$

Finally, the interaction energy is the sum of the attraction energy between all the units of the top parts and the under parts of the liquid. If the sectional area of the liquid is a , the energy is described as:

$$U_4 = \int_{r_0}^{\infty} \frac{\pi N^2 A}{6r^3} d_x = \frac{\pi N^2 A}{12r_0^2} a$$

In this process, the increase surface area is $2a$, and the increase of unit energy value is described as

$$U^s = \frac{\pi N^2 A}{12r_0^2}$$

This is an ideal result and is applicable while the liquid is in vacuum. The fact is that when the *surface* is mentioned, it must be referred to as that; the surface connects with steam. So U^s can be written as:

$$U^s = \frac{\pi(N_L - N_V)^2 A}{12r_L^2}$$

N_L and N_V are the molecule number of liquid and steam phase per unit, and r_L is the distance between balance molecules (Adam 1941; Moore 1962; White 1948).

As mentioned, an important characteristic of a liquid penetrating material is its ability to freely wet the surface of an object (Woodruff 2002). At the liquid-solid interface, if the molecules of the liquid have a stronger attraction to the molecules of the solid surface than to each other (the adhesive forces are stronger than the cohesive forces), wetting of the surface occurs. Alternately, if the liquid molecules are more strongly attracted to each other than to the molecules of the solid surface (the cohesive forces are stronger than the adhesive forces), the liquid beads up and does not wet the surface of the part.

Young's Equation

Young's equation defines the balance of forces caused by a wet drop on a dry surface (Tadmor 2004). If the surface is hydrophobic then the contact angle of a drop of water will be larger. Hydrophilicity is indicated by smaller contact angles and higher surface energy. (Water has rather high surface energy by nature; it is polar and forms hydrogen bonds.) Young's equation gives the following relation (de Gennes 1985):

$$\gamma_{SL} + \gamma_{LV} \cos \theta_c = \gamma_{SV} x$$

where γ_{SL} , γ_{LV} , and γ_{SV} are the interfacial tensions between the solid and the liquid, the liquid and the vapor, and the solid and the vapor, respectively. The equilibrium contact angle that the drop makes with the surface is denoted by θ_c . To derive Young's equation, normally the interfacial tensions are described as forces per unit length, and from the one-dimensional force balance along the x -axis Young's equation is obtained.

Young's equation assumes a perfect flat surface, and in many cases surface roughness and impurities cause a deviation in the equilibrium contact angle from the contact angle predicted by Young's equation. Even on a perfectly smooth surface a drop will assume a wide

spectrum of contact angles, ranging from the advancing contact angle, θ_A , to the receding contact angle, θ_R . The equilibrium contact angle (θ_c) can be calculated from θ_A and θ_R , as was shown by (Tadmor 2004):

$$\theta_c = \arccos\left(\frac{r_A \cos \theta_A + r_R \cos \theta_R}{r_A + r_R}\right)$$

Where

$$r_A = \left(\frac{\sin^3 \theta_A}{2 - 3 \cos \theta_A + \cos^3 \theta_A}\right)^{\frac{1}{3}};$$

$$r_R = \left(\frac{\sin^3 \theta_R}{2 - 3 \cos \theta_R + \cos^3 \theta_R}\right)^{\frac{1}{3}}$$

In the case of dry wetting, one can use the Young-Dupré equation, which is expressed by the work of adhesion. This method accounts for the surface pressure of the liquid vapor, which can be significant. Pierre-Gilles de Gennes, a Nobel Prize laureate in physics, describes wet and dry wetting and how the difference between the two relates to whether or not the vapor is saturated.

Contact Angle

Another way to quantify a liquid's surface wetting characteristics is to measure the contact angle of a drop of liquid placed on the surface of an object. The contact angle is the angle formed by the solid/liquid interface and the liquid/vapor interface measured from the side of the liquid (Fig. 3). Liquids wet surfaces when the contact angle is less than 90° . For a penetrant material to be effective, the contact angle should be as small as possible. In fact, the contact angle for most liquid penetrants is very close to 0° .

Wetting ability of a liquid is a function of the surface energies of the solid-gas interface, the liquid-gas interface, and the solid-liquid interface. The surface energy across an interface or the surface tension at the interface is a measure of the energy required to form a unit area of new surface at the interface. The intermolecular bonds or cohesive forces between the molecules of a liquid cause surface tension. When the liquid encounters another substance, there is

usually an attraction between the two materials. The adhesive forces between the liquid and the second substance will compete against the cohesive forces of the liquid. Liquids with weak cohesive bonds and a strong attraction to another material (or the desire to create adhesive bonds) will tend to spread over the material. Liquids with strong cohesive bonds and weaker adhesive forces will tend to bead-up or form a droplet when in contact with another material.

In liquid penetrant testing, there are usually three interfaces involved, the solid-gas interface, the liquid-gas interface, and the solid-liquid interface. For a liquid to spread over the surface of a part, two conditions must be met. First, the surface energy of the solid-gas interface must be greater than the combined surface energies of the liquid-gas and the solid-liquid interfaces. Second, the surface energy of the solid-gas interface must exceed the surface energy of the solid-liquid interface.

A penetrant's wetting characteristics are also largely responsible for its ability to fill a void. Penetrant materials are often pulled into surface breaking defects by capillary action. The capillary force driving the penetrant into the crack is a function of the surface tension of the liquid-gas interface, the contact angle, and the size of the defect opening. The driving force for the capillary action can be expressed as the following formula:

$$\gamma = \frac{Pl}{\pi r(l - 2r)}$$

where:

r = radius of the crack opening ($2pr$ is the line of contact between the liquid and the solid tubular surface.)

σ_{LG} = liquid-gas surface tension

θ = contact angle

Five Approaches for Determining the Energy of Solid Surfaces

Antonoff's Rule

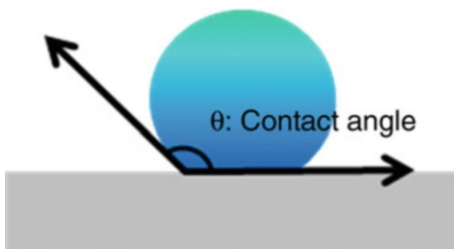
Antonoff thought that surface tension γ_{ab} between two kinds of liquids saturated by each other is the difference between the surface tension of the two kinds liquid:

$$\gamma_{ab} = |\gamma_a - \gamma_b|$$

γ_a : the surface tension of a , which is saturated by b

γ_b : the surface tension of b , which is saturated by a

Moreover, experiments proved that the Antonoff rule has some limitations and is not applicable for all systems.



Surface Free Energy, Fig. 3 Contact angle

Good-Girifalco Theory

Good and Girifalco took into account the force among the molecules that changes with the change of molecules' characterization, and they thought that the surface energy decreases when liquid *a* adheres to *b*. The equation is as follows:

$$W_a = \gamma_a + \gamma_b - \gamma_{ab}$$

Good and Girifalco pointed out that, due to the differences in volume and molecules interior force between two kinds molecules, the equation should corrected as follows:

$$\gamma_{ab} = \gamma_a + \gamma_b - 2\Phi\sqrt{\gamma_a\gamma_b}$$

$$\Phi = \frac{\gamma_a + \gamma_b - \gamma_{ab}}{2\sqrt{\gamma_a\gamma_b}}$$

Fowkes' Theory

This approach divides the surface energy into two components, dispersive and polar, and uses a geometric mean approach to combine their contributions. The resulting equation when combined with Young's equation yields:

$$\gamma = \gamma^d + \gamma^p$$

where γ^d is the dispersive force and γ^p is polar force. The superscripts *d* and *p* refer to the dispersive and polar components of each. If there is only dispersive force between the two different kinds of molecules, the equation can be changed as follows:

$$\gamma_{ab} = \gamma_a + \gamma_b - 2(\gamma_a^d\gamma_b^d)^{\frac{1}{2}}$$

Harmonic Mean (Wu)

This method utilizes a similar approach but uses a harmonic mean equation to sum the dispersive and polar contributions. Contact angles against two liquids with known values of C_{ab} and C_{bb} are measured. The values for each experiment are put into the following equation:

$$C_{ab} = \frac{3}{4}h\nu_a\alpha_a^2$$

$$C_{bb} = \frac{3}{4}h\nu_b\alpha_b^2$$

where *h* refers to Planck's constant, α refers to molecule polar component, and ν refers to molecular characteristic vibrational frequency (Vicente 2002).

When $\nu_a = \nu_b$:

$$C_{ab} = (C_{aa}C_{bb})^{\frac{1}{2}}$$

When $\alpha_a^2 = \alpha_b^2$:

$$C_{ab} = \frac{2C_{ab}C_{bb}}{C_{ab} + C_{bb}}$$

Acid-Alkali Theory

Fowkes' study indicated that the above-mentioned surface tension polar part equations are not accurate enough. Because the polar forces that can be treated with geometrical average or reciprocal average are only Keesom force and Debye force. However, their contribution to polar force is small. For the main part, which is important in γ^b , this average treatment is not applicable. Fowkes pointed out that if two kinds of liquids are both Lewis acid or alkali, the spanned surface force only needs to take into account dispersive force (geometrical average); if one is Lewis acid and the other is Lewis alkali, one needs not only to take into account dispersive force but also electron transfer action. Then, the equation is given by

$$\gamma_{ab} = \gamma_a + \gamma_b - w_a^d - w_a^{AB} - w_a^p$$

$$w_a^{AB} = N_{ab} + \varepsilon_{AB} = \frac{1}{a} \left(-\frac{\Delta H_{AB}}{N_0} \right)$$

$$w_a^d = 2(\gamma_a^d\gamma_b^d)^{1/2}$$

w_a^d : the contribution of stretching work when only dispersive force exists

w_a^{AB} : the contribution to spread work of acid-alkali action

N_{ab} : the molecule pair number of the two kinds liquid on unit area

ε_{AB} : the acid-alkali action energy of each molecule pair

a: the area of each molecule pair

N_0 : Avogadro constant.

The surface energy of a liquid may be measured by stretching a liquid membrane (which increases the surface area and hence the surface energy density). However, such a method cannot be used to measure the surface energy of a solid because stretching of a solid membrane induces elastic energy in the bulk in addition to increasing the surface energy.

The surface energy of a solid is usually measured at high temperatures. At such temperatures the solid creeps and, even though the surface area changes, the volume remains approximately constant. If γ is the surface energy density of a cylindrical rod with radius *r* and length *l* at high temperature and a constant uniaxial tension *P*, then at equilibrium, the variation of the total Gibbs free energy vanishes and it can be drawn out that

$$\delta G = -P\delta l + \gamma\delta A = 0 \Rightarrow \gamma = P \frac{\delta l}{\delta A}$$

where G is the Gibbs free energy and A is the surface area of the rod:

$$A = 2\pi r^2 + 2\pi rl \Rightarrow \delta A = 4\pi r\delta r + 2\pi l\delta r + 2\pi r\delta l$$

Also, since the volume (V) of the rod remains constant, the variation (δV) of the volume is zero, that is,

$$\begin{aligned} V = 2\pi r^2 l = \text{const} \tan t \Rightarrow \delta V \\ = 2\pi r l \delta r + 2\pi r^2 \delta l = 0 \Rightarrow \delta r = -\frac{r}{2l} \delta l \end{aligned}$$

Therefore, the surface energy density can be expressed as

$$\gamma = \frac{Pl}{\pi r(l - 2r)}$$

The surface energy density of the solid can be computed by measuring P , r , and l at equilibrium.

Cross-References

- [Contact Angle](#)
- [Interface](#)
- [Surface Energy](#)
- [Surface Tension](#)
- [Wetting Ability](#)

References

- N.K. Adam, *The Physics and Chemistry of Surfaces*, 3rd edn. (Oxford University Press, London, 1941)
- J.B. Calvert, *Surface Tension*, University of Denver. Dataphysics. August 8 (2007)
- P.G. de Gennes, Wetting: statics and dynamics. *Reviews of Modern Physics* **57**(3), 827–863 (1985)
- P.-G. de Gennes, F. Brochard-Wyart, D. Quéré, A. Reisinger, *Capillary and Wetting Phenomena – Drops, Bubbles, Pearls, Waves* (Springer, New York, 2002)
- C. Emil, Surface free energy of sulfur – revisited I. Yellow and orange samples solidified against glass surface. *Journal of Colloid and Interface Science* **319**, 505 (2008)
- R. Förch, H. Schönherr, A.T.A. Jenkins, *Surface Design: Applications in Bioscience and Nanotechnology* (Wiley-VCH, Weinheim/New York, 2009), 471
- W.J. Moore, *Physical Chemistry*, 3rd edn. (Prentice Hall, Englewood Cliffs, 1962)
- H.E. White, *Modern College Physics* (van Nostrand, New York, 1948)
- D.P. Woodruff, *The Chemical Physics of Solid Surfaces* (Elsevier, Amsterdam, 2002)
- Wikipedia. (May 10, 2012) Wetting: http://en.wikipedia.org/wiki/Surface_energy (May 10, 2012).
- R. Tadmor, Line energy and the relation between advancing, receding and young contact angles. *Langmuir* **20**(18), 7659 (2004)
- Wikipedia. (May 9, 2012) Contact angle: http://en.wikipedia.org/wiki/Contact_angle (May 9, 2012).

Wikipedia. (May 7, 2012) Surface tension: http://en.wikipedia.org/wiki/Surface_tension (May 7, 2012).

Wikipedia. (April 10, 2012) Surface energy: http://en.wikipedia.org/wiki/Surface_energy (April 10, 2012).

C. Vicente, W. Yao, H. Maris, G. Seidel, Surface tension of liquid 4He as measured using the vibration modes of a levitated drop. *Physical Review B* **66**(21) (2002)

Surface Hardening

► [Induction Heat Treating](#)

Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH

ULRICH BAUDIS¹, HERVÉ CHAVANNE^{2,3}, PHILIPPE MAURIN-PERRIER²

¹Durferit GmbH, Mannheim, Germany

²HEF R&D, Z.I. Sud, Andrézieux–Bouthéon Cedex, Germany

³H.E.F. USA, Springfield, OH, USA

Synonyms

[Ionic liquids/molten salt stainless steel hardening](#)

Definition

A low-temperature carburizing process applicable on austenitic stainless steel. The treatment increases the wear resistance and surface hardness by implementation of carbon (mainly) and nitrogen (minor) in solid solution into the metallic surface. Formation of chromium carbides and nitrides is suppressed and loss of corrosion resistance is avoided.

Scientific Fundamentals

General Description

Stainless steel is characterized by a minimum content of 10.5 weight % of chromium in the iron matrix. A higher chromium content and additional alloying elements – especially nickel, molybdenum, niobium, and titanium – improve the corrosion resistance against various corrosive media like humid air, industrial atmospheres, diluted acids, sulfides, salt water, and so on ([Merkblatt 821](#); Gümpel et al. 1996). The manifold types

of stainless steel are divided into four groups according to their lattice structure at room temperature:

1. Austenitic stainless steel
2. Ferritic stainless steel
3. Ferritic-austenitic (“duplex”) stainless steel
4. Martensitic stainless steel

Pure iron exists in two lattice configurations: the cubic body-centered structure (α -iron, ferrite) and the cubic face centered structure (γ -iron, austenite, above 911 °C). The same lattice structures are found in stainless steel, however, the range of existence of the structures is different. Several Cr-Ni-stainless steel alloys (Cr content approx. 18 wt.%, Ni > 8 wt.%) exist at room temperature in an austenitic structure. Austenite of pure iron (cubic face centered lattice) has a good solubility for carbon up to 2.1 wt.%. Accordingly, austenitic stainless steels are able to take up similar amounts of carbon in solid solution as they have the same structure at much lower temperatures.

Austenitic stainless steels can be polished to mirror shine and resist humid atmospheres and many corrosive media very well. A significant disadvantage of austenitic stainless steel is its low hardness (200–300 HV 0.1), which cannot be increased by thermal treatment. The tribological problem is mainly adhesive wear in the form of seizure (i.e., scuffing and galling).

If austenitic stainless steel is surface hardened by conventional methods like nitrocarburizing (580 °C) or boriding (>900 °C) the surface hardness will be increased and certain wear problems might be solved, but after such treatment the corrosion resistance on the surface will be lost completely. The reason is that chromium is converted into chromium carbide, chromium nitride, or chromium boride and is thus withdrawn from the metallic matrix and consequently the surface area is no longer a stainless alloy (Table 1).

Description of the Durofer® SH Process

The Durofer® SH process can be simply characterized as a “low temperature carburizing salt bath treatment.”

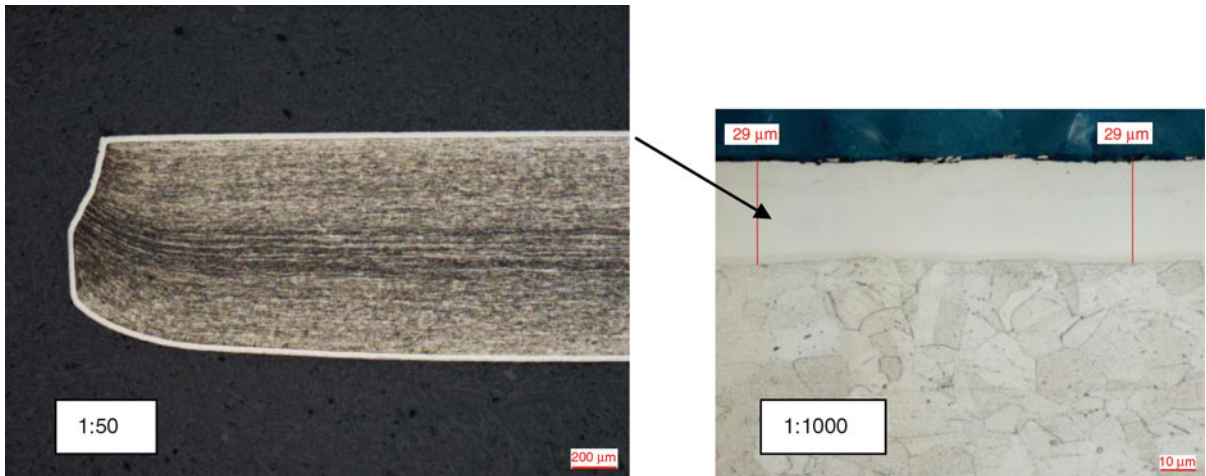
The Durofer SH process carbon (and to some extent nitrogen) are implemented by diffusion into the surface of the stainless austenitic component at such low temperatures (350–420 °C) that these elements remain in solid solution within the austenitic lattice and do not form chromium carbides or nitrides. Thus, the metallic matrix is not affected by withdrawal of chromium, the alloy is still a stainless composition, and the corrosion resistance of the base material is substantially preserved. Process control involves chemical analysis of the active salt components, and control of temperature and treatment time, which is usually 24 h. After treatment carbon is found in a concentration of 2–3 weight % and nitrogen up to 12–14 %, both in solid solution, within a surface layer of 5–30 μm thickness. The presence of these elements in the diffusion layer leads to high compressive strength and as a consequence to a hardness of 700–1,200 HV 0.05 or HK 0.025. The carbon and nitrogen containing austenite thus obtained is designated as “expanded austenite” or “S-phase” austenite in the literature (Christiansen and Somers 2006; Bell and Sun 1999).

Figures 1 and 2 show the cross section of a typical layer obtained by the Durofer® SH process on stainless steel 1.4301 and 1.4401 (AISI 316) after 48 h treatment. The expanded austenite layer is clearly visible in contrast to the base material after etching with stainless steel etchant or Marble’s etchant. The outer part of the layer in Fig. 2 (0–7 μm) is slightly darker and indicates the presence of nitrogen together with carbon in solid solution in this area. The total diffusion depth is approximately 20 μm . The appearance of the layer in Fig. 2 corresponds well with the concentration profiles in Fig. 3. The layer thickness depends mainly on the duration of the treatment.

Concentration profiles of carbon and nitrogen within the layer can be taken from GDS spectra (glow discharge optical emission spectrometry). Figure 3 shows a typical GDS spectrum of a Durofer® SH-treated surface. Carbon is distributed throughout the layer whereas nitrogen is accumulated only in the outer sphere of the layer. Oxygen is found in traces only on top of the layer.

Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Table 1 Common austenitic stainless steel

UNI	EN	C max.	Cr	Ni	Mo	Nb	Ti
X5CrNi18-10	1.4301	0.07	17.0–19.5	8.0–10.5			
X6CrNiTi18-10	1.4541	0.08	17.0–19.0	9.0–12.0		< 1	<0.7
X2CrNiMo17-12-2	1.4404	0.03	16.5–18.5	10.0–13.0	2–2.5		
X6CrNiMoTi17-12-2	1.4571	0.08	16.5–18.5	10.5–13.5	2–2.5		<0.7
X2CrNiMo18-14-3	1.4435	0.03	17.0–19.0	12.5–15.0	2.5–3		(w.- %)



Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 1 Typical layer on 1.4301 (Durofer SH/48 h/415 °C) ~30 μm. Etchant: Marble's



Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 2 Cross section 1.4401 after Durofer SH 48 h/400 °C/ Marble's/1:1,000

Figure 4 shows some typical micro hardness curves of Durofer® SH-treated stainless samples. The hardness curves are in compliance with the cross sections.

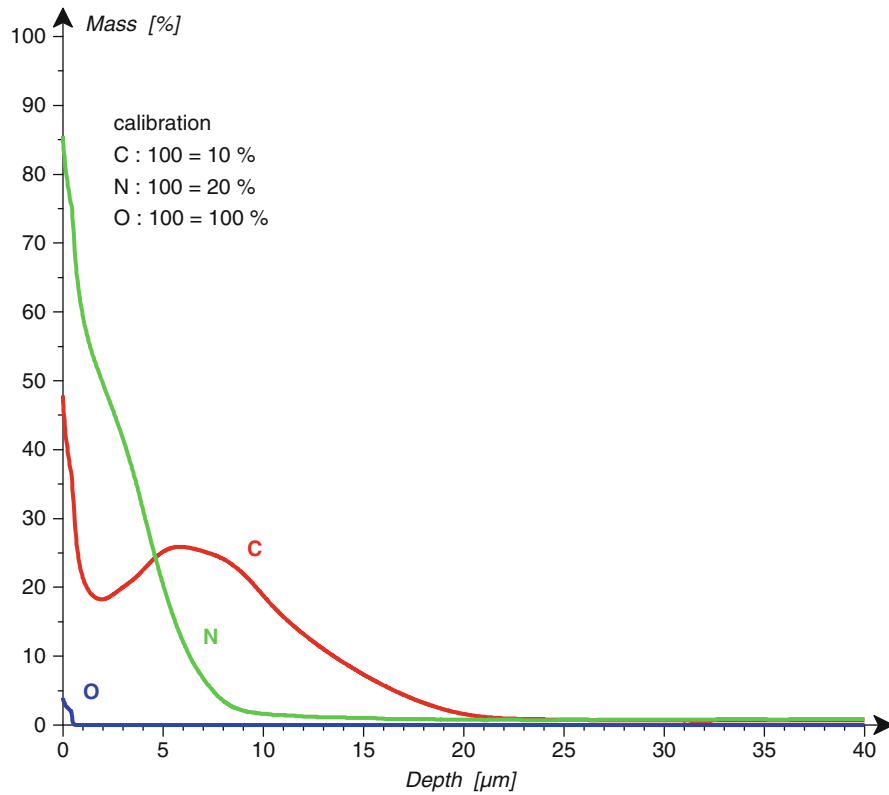
Component Properties after Durofer® SH Treatment

The Durofer® SH treatment forms a hard yet ductile layer on the surface, which accrues from the base material and is

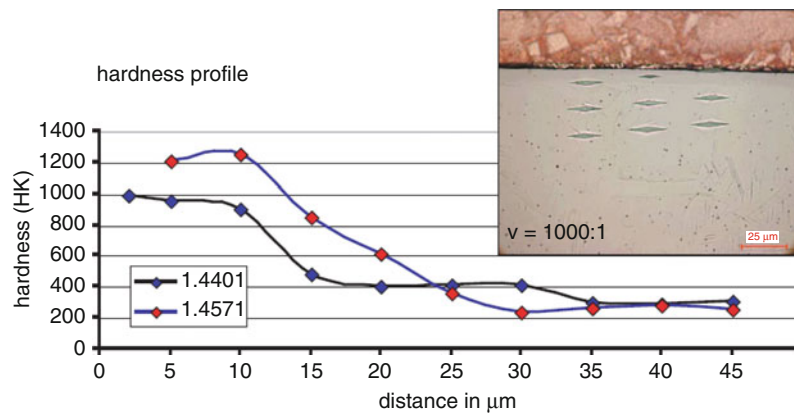
firmly fixed to it, unlike galvanic layers, which are just laid upon. Corrosion creep or surface flaking is thus impossible. The layer follows the contours of the workpiece perfectly and shows a superb uniformity (Fig. 1). Durofer® SH increases the hardness of the steel surface significantly (Fig. 4). Bending or rotating fatigue strength is increased by 20–30 % above its original value due to the formation of compressive strength on the surface (Fig. 5).

GDS-conditions:**1200V, 6mA / 0,0 hPa**

date / time :

14.01.2009 10:16:40

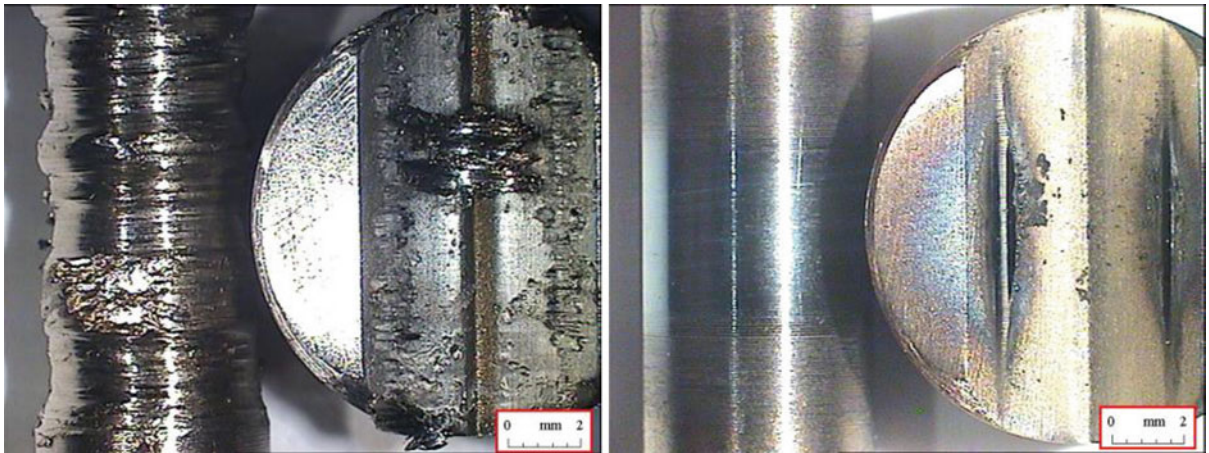
Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 3 GDOES (1.4401) Durofer SH 48 h/400 °C



Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 4 Hardness profiles after Durofer SH treatment



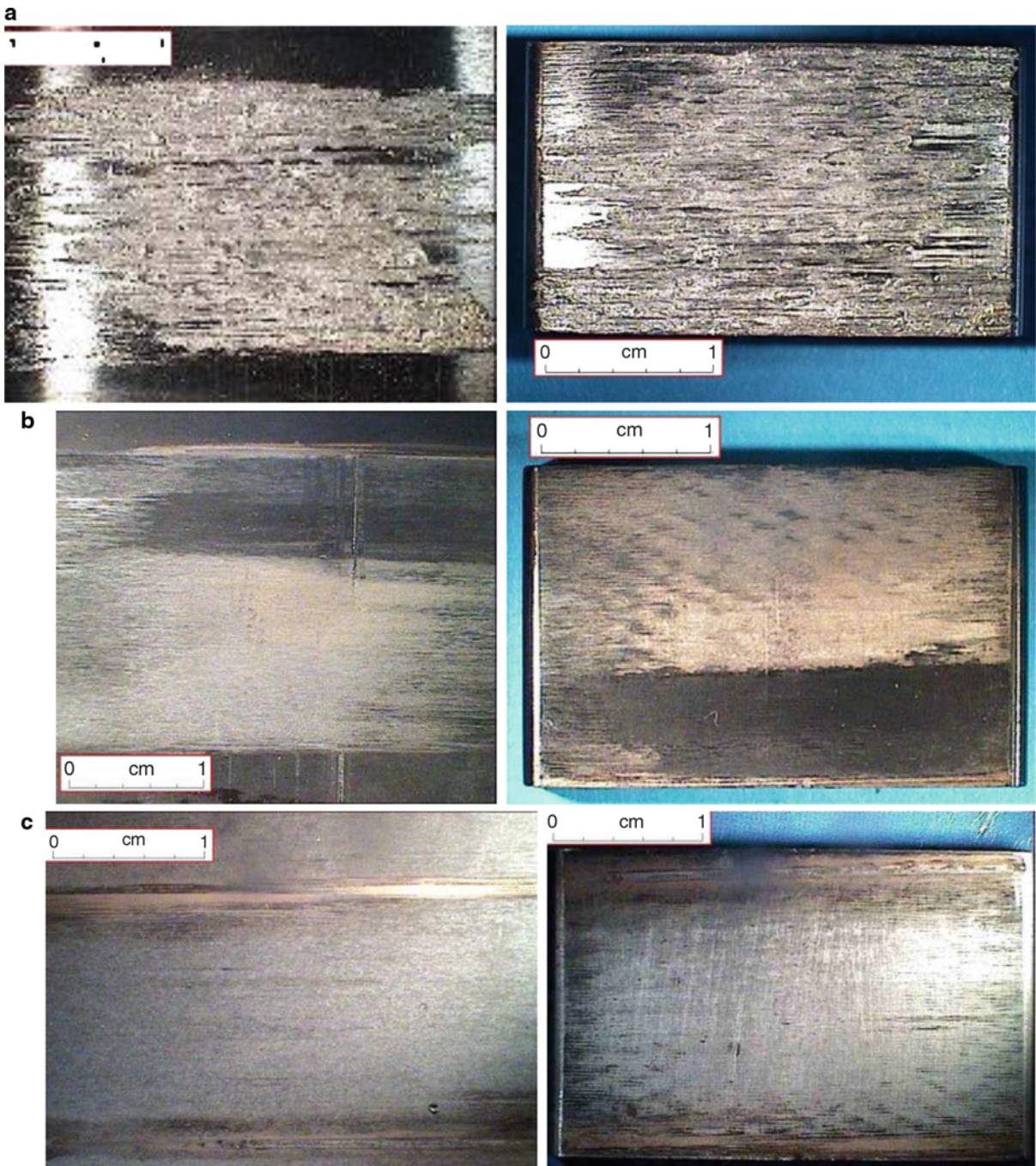
Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 5 High cycles fatigue strength : rotating bending probes – non-carved/X6CrNiTi18-10 (1.4541). Treatment: Durofer SH/12 h/layer thickness 8 μ m



Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 6 Faville test results. (a) Not treated stainless steel 1.4301. Test start at 100 DaN. 2 s – 100 DaN test stop. Seizing. (b) Durofer SH treated/12 h/400 °C. Test start at 100 DaN. 0 s – 100 DaN. 10 s – 300 DaN. 12 s – 350 DaN test stop

Treated parts have good resistance against sliding wear. Scuffing and galling are reduced remarkably (Figs. 6, 7). The surface becomes resistant against scratches caused by smut or fine dust particles. The color of treated parts

changes from silvery to a light gray appearance, but the surface can still be polished to a shiny finish. The good corrosion resistance of the stainless material is preserved, however, there are some limitations, described below



Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 7 Adhesive wear test/"flat-flat-test"/stainless steel 1.4301. left side: track – right side: pin. test conditions: speed 0,01 m/s – length 0,01 m – cycle time 2 s – test duration 35 min. **(a)** 1.4301 not treated, load 100 DaN, seizure. **(b)** Durofer SH treated, load 100 DaN, no seizure. **(c)** Durofer SH treated, load 200 DaN, increased load, no seizure

under “Limits” and “Preconditions.” The common “salt spray test” (according to ASTM B 117 or DIN EN ISO 9227) may not be the best test method, as stainless austenitic steels are sensitive to pit corrosion in chloride media anyway. A better choice for testing corrosion properties is the measurement of potentiodynamic anodic current density curves before and after treatment.

The significantly improved resistance of treated stainless steel against adhesive wear is impressively demonstrated by the Faville test (Fig. 6) and the “flat/flat test” (Fig. 7).

Preconditions

Austenitic stainless steel of minor quality may contain some amount of delta-ferrite, which can be identified by a strong magnet; its level in the base material should be set to zero when ordering material from the steel manufacturer. Cold-working, severe turning, or cold extrusion may cause formation of martensite on the surface of stainless austenitic steel. Both must be avoided, otherwise the corrosion resistance after Durofer® SH treatment might not be preserved sufficiently. Martensite can be eliminated, usually by solution annealing in a high vacuum prior to the Durofer® SH treatment.

Limits

Ferritic and martensitic stainless steel cannot be treated as they do not have the necessary austenitic (cubic face centered) lattice structure. Duplex steel can be treated, but results regarding the preservation of the corrosion resistance should be tested case by case. Treated parts must not be exposed to temperatures above 350 °C for a long time, as carbon and nitrogen would start to react with chromium and the other alloying elements to form carbides and nitrides and would be withdrawn from the

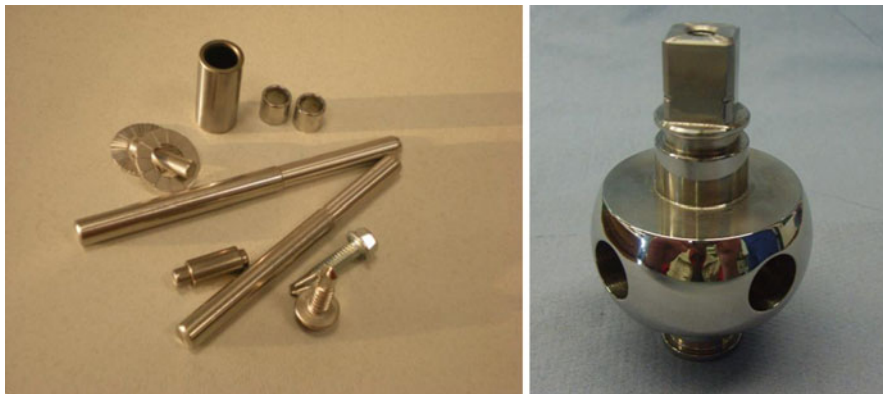
metallic matrix, hardness and corrosion resistance would be lost.

Compatible Processes

Few processes are known that may be compatible with the Durofer® SH process. An early attempt is the Kolsterising® process, developed in the Netherlands during the 1980s (Van der Jagt 2000; Rey and Jacquot 2002). Process details are not published or patented, apart from that it is said to be a gaseous treatment (Rey and Jacquot 2002). Other processes are based on the diffusion of carbon by plasma carburizing (Günther et al. 2001) or gas carburizing (Aoki and Kitano 2002) or nitrogen by plasma nitriding into the surface of stainless steel at low temperatures (Christiansen and Somers 2006). These processes suffer from necessary costly depassivation prior to the treatment and a rather sophisticated process equipment and control.

Key Applications

Austenitic stainless steels combine very good corrosion resistance, mechanical properties, and workability (Merkblatt 821) and therefore are widely used as construction material in the food industry, electrical industry, pharmaceutical and medical technology, hydraulic and pneumatic devices, chemical process engineering, the offshore industry, and related industries. For many devices used in these fields of technology Durofer® SH can provide improved performance. Pistons, tubes, fittings, washers, valves, pin rods, and control rods as well as pins and bushings of stainless steel chains are typical applications among a variety of others. Figure 8 shows a small choice of potential applications. As an example, the sliding behavior of pin rods in gas springs used in nautical applications and pneumatic cylinders is much



Surface Hardening of Austenitic Stainless Steel/Durofer(R) SH, Fig. 8 Potential applications for Durofer® SH treatment

improved; the same is true for the performance of ball valves, metering, and gate valves. Adjusting rods and pump impellers are also potential applications as well as chain pins and bushings.

Cross-References

- [Carburizing and Carbonitriding](#)
- [Corrosive Wear](#)
- [Fatigue](#)
- [Fatigue Limit](#)
- [Friction Coefficient](#)
- [Surface Roughness](#)

References

- K. Aoki, K. Kitano, Surface hardening for austenitic stainless steels based on carbon solid solution. *Surf. Eng.* **18**, 462–464 (2002)
- T. Bell, Y. Sun et al., The response of austenitic stainless steel to low-temperature plasma nitriding. *Heat Treat. Met.* **1**, 9–16 (1999)
- Th. Christiansen, M.A.J. Somers, Mikrostrukturausbildung beim Randschichthärten von rostfreiem Stahl im Gas bei niedrigen Temperaturen, *Struers Zeitschrift für Materialografie* **9** (2006), pp. 1–17, (supported by TU Danmark, Lyngby)
- V. der Jagt, Kolsterising – surface hardening of austenitic and duplex stainless steel without loss of corrosion resistance. *Heat Treat. Met.* **3**, 62–65 (2000)
- P. Gümpel et al., *Rostfreie Stähle*, 3rd edn, Expert-Verlag, Renningen-Malmsheim, 1996, ISBN 3-8169-1735-6
- H. Günther, Mayr, Jung, Oberflächenhärtung von austenitischen Stählen unter Beibehaltung der Korrosionsbeständigkeit. *HTM - Härtereitechnische Mitteilungen* **56**, 74–83 (2001)
- Merkblatt 821 - Edelstahl, edited by Informationsstelle Edelstahl Rostfrei, Post Box 10 22 05 D-40013, Düsseldorf
- O. Rey, P. Jacquot, Kolsterising: Hardening of austenitic stainless steel. *Surf. Eng.* **18**, 412–414 (2002)

Surface Indentation and Failure

- [Failure Mechanisms of Rolling Element Bearings](#)

Surface Integrity

- [Tribological Effects of Machining Carbon Nanotube Composites](#)

Surface Morphology

- [Topography of Engineering Surfaces](#)

Surface Nanocrystallization and Hardening (SNH)

LEON L. SHAW

Department of Chemical, Materials and Biomolecular Engineering, Institute of Materials Science, University of Connecticut, Storrs, CT, USA

Synonyms

S^2PD – Surface severe plastic deformation; SMAT – Surface mechanical attrition treatment; USSP – Ultrasonic shot peening

Definition

Surface nanocrystallization and hardening (SNH) is a process that relies on surface severe plastic deformation (S^2PD) induced by repeated impacts of high-energy balls to produce a nanocrystalline and hardened surface layer. This process is similar to conventional shot peening (SP) in the sense that both processes entail repeated impacts of the workpiece surface by high-speed balls or shots. However, SNH offers higher impact energies than SP (e.g., 50–180 times larger) because balls of 3–8 mm in diameter are used in SNH and shots of 0.2–0.3 mm are typically used in SP (Dai and Shaw 2007). As a result of such a large difference in the impact energy, SNH always leads to surface nanocrystallization, whereas SP does not. The high velocity of balls in SNH can be generated through collision between balls and a vibrating chamber driven by an ultrasonic generator or by an electric motor (Dai and Shaw 2007; Wu et al. 2002). Alternatively, the high velocity can be created through a high-pressure light-gas gun that accelerates particles to a desired impact speed (Umemoto et al. 2004). Different devices invented for generating high-speed balls offer a wide range of kinetic energies to produce various degrees of surface severe plastic deformation.

Scientific Fundamentals

Surface Deformation Zone

When the surface of a workpiece is subjected to repeated impacts of high-energy balls (such as steel or WC balls of 3–8 mm with a speed of 5–15 m/s), the surface region of the workpiece will undergo several changes simultaneously. These include (a) formation of a work-hardened layer, (b) introduction of residual compressive stresses, (c) formation of a nanocrystalline (nc) surface layer, (d) increased surface roughness, and (e) sometimes

surface contamination owing to material transfer between the workpiece and balls (Dai and Shaw 2007; Ortiz et al. 2008). All of these changes are caused by surface severe plastic deformation. The extent of surface severe plastic deformation, the depth of the plastic deformation zone, the residual stress profile, and the surface roughness can all be simulated through finite element modeling of dynamic loading (Dai and Shaw 2007; Dai et al. 2004a). Alternatively, the depth of plastic zone and the surface roughness induced by SNH can be estimated via analytical approaches (Al-Obaïd 1995; Shaw and DeSalvo 1970; Dai et al. 2004b) if the following two rules are allowed:

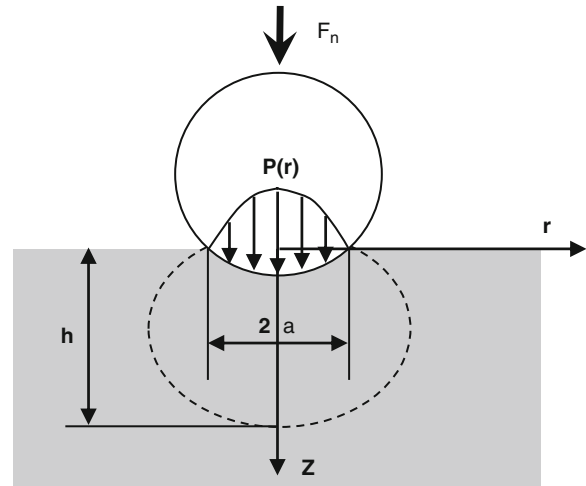
1. The stress field generated by elastic impact (with moderate velocities) is identical to that generated by elastic contact.
2. The normal force upon impact is due to the deceleration of the moving ball.

With these two rules, dynamic loading can be approximated by static contact, and the depth of plastic zone, h , can be related to the parameters of SNH and SP processes via the following equation (Al-Obaïd 1995):

$$\frac{h}{R} = 3 \left(\frac{2}{3} \right)^{\frac{1}{4}} \left(\frac{\rho v_b}{\bar{P}} \right)^{\frac{1}{4}} \quad (1)$$

where \bar{P} is the mean normal pressure in the circle of contact (see Fig. 1), R the radius of the ball, v_b the impact velocity of the ball, and ρ the ball density. The depth of plastic zone estimated from (1) does not include work hardening, strain rate sensitivity, and thermal effects. Nevertheless, (1) has been shown to be adequate in estimating the depth of plastic zone when compared with experiments (Al-Obaïd 1995; Shaw and DeSalvo 1970). Thus, (1) offers a convenient way to guide the selection of the process parameters of SNH in order to achieve a certain depth of the plastic zone. As shown in (1), the depth of plastic zone increases with the ball size, ball velocity, and ball density. Since the mean normal pressure is proportional to the hardness of the workpiece, (1) also reveals that the depth of plastic zone decreases with increasing the hardness of the workpiece.

The depth of plastic zone is one of the most important parameters in SNH because it determines the thickness of the work-hardened layer as well as the thickness of the surface layer within which residual compressive stresses are present. Finite element modeling has shown that the depth of plastic zone coincides with the thickness of the surface layer with residual compressive stresses (Meguid et al. 1999). The depth of plastic zone also has the direct influence on the thickness of the nc surface layer because experiments have revealed that exceeding a critical plastic

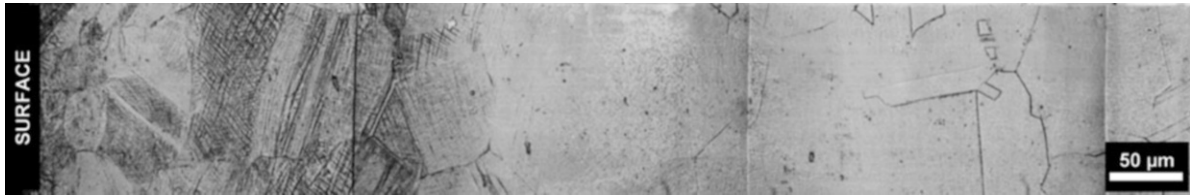


Surface Nanocrystallization and Hardening (SNH), Fig. 1 Schematic of the distribution of Hertz pressure during ball impact; the area within the *dashed line* represents the plastic zone in the deformed solid. Here, F_n is the maximum normal force upon impact, which is related to the deceleration of the impacting ball. $P(r)$ is the profile of the normal pressure in the circle of contact and can be used to calculate the mean normal pressure, \bar{P} . $2a$ is the diameter of the contact circle, and h is the depth of plastic zone

strain is necessary for the formation of nano-grains in the severe plastic deformation process (Umemoto et al. 2004). A deeper plastic zone means a thicker surface region that has the effective plastic strain exceeding the critical strain and thus a thicker nc surface layer. This is the reason why SNH always leads to surface nanocrystallization, whereas SP does not.

Microstructure Gradient

Since severe plastic deformation only takes place at the surface region in the SNH process, it would be expected that a microstructure gradient with a nc surface and a coarse-grained interior will be generated by SNH. This is indeed the case, as uncovered by many experimental studies (Wu et al. 2002; Ortiz et al. 2008; Tao et al. 2002). Figure 2 shows an optical micrograph of the cross-section of a C-2000 alloy treated with SNH for 70 min using WC/Co balls of 7.9 mm in diameter. The C-2000 alloy is a single phase material with a face-centered-cubic (FCC) crystal structure and the following chemical composition (wt%): 23Cr, 16Mo, 1.6Cu, 0.01 C, 0.08Si, and balance Ni. Several microstructural features are noted from Fig. 2. (a) Grains in the annealed material – those far from the impacted surface – are approximately equiaxed in shape,



Surface Nanocrystallization and Hardening (SNH), Fig. 2 The cross-sectional optical microstructure of an annealed C-2000 alloy after SNH-processing for 70 min using WC/Co balls of 7.9 mm in diameter

50–100 μm in diameter, with fairly straight grain boundaries and possess annealing twins, 5–20 μm wide. (b) Many deformation markings (i.e., straight and bent lines) are visible near the processed surface. These markings are confined within individual grains and their densities increase as the distance to the processed surface diminishes. These deformation markings have been identified to be deformation twinning (Villegas et al. 2005; Shaw et al. 2008). In the proximity to the surface, these twins populate entire grains, suggesting that the intersection of the twins contributes to structural refinement. (c) Careful observation of the specimen reveals that the greatest depth at which deformation twins are present is approximately 825 μm , indicating that the depth of plastic zone for this particular processing condition is on the order of 825 μm . (d) The microstructure at and directly underneath the surface is difficult to resolve with an optical instrument. The twins and intersections, clear at greater depths, become barely identifiable near the surface of the specimen, indicating the greater structural refinement at the surface than in the sub-surface region.

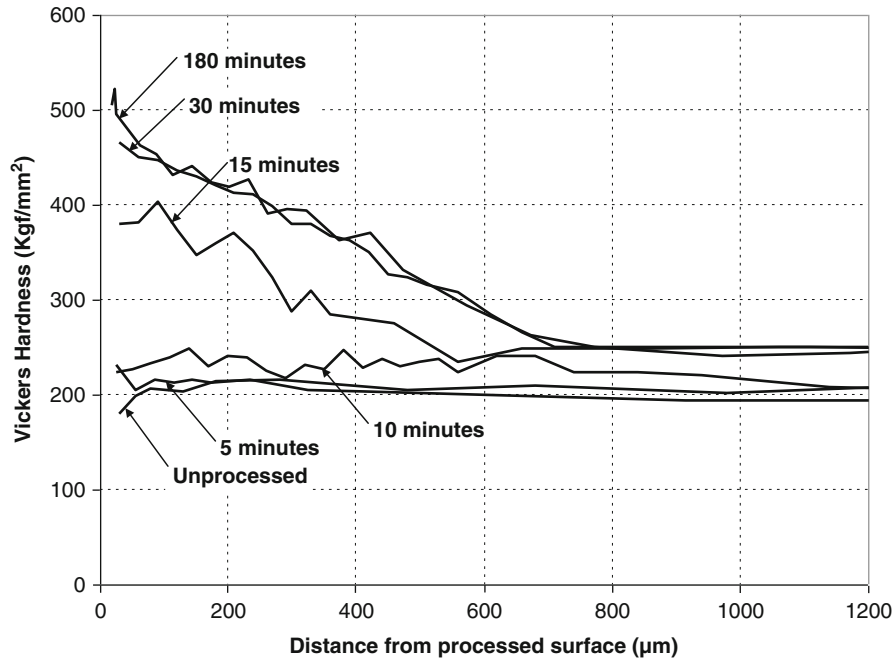
In addition to the microstructure gradient observed using optical and scanning electron microscopy (SEM), the gradient in grain sizes determined using transmission electron microscopy (TEM) and X-ray diffraction (XRD) has also been established. With the aid of TEM, it has been shown that the grain size at the impacted surface is 7 nm for a commercially pure Fe after the surface treatment using stainless steel balls of 8 mm in diameter for 60 min (Tao et al. 2002). This grain size increases to ~ 100 and 1,000 nm at 15 and 40 μm away from the impacted surface, respectively. At the depth of 110 μm no change in the grain size is observed (Tao et al. 2002). A similar grain size gradient is observed for an Al alloy 7075 treated using a USSP device for 15 min (Wu et al. 2002). It is found that the thickness of the nc surface layer (grain size < 100 nm) is about 20 μm , which is followed by a submicro-grained region (0.1–1 μm) of 35 μm thick. Next to this submicro-grained region is an extended microband of ~ 10 μm thick, characterized by elongated

grains (> 1 μm). The total thickness of the deformation zone is ~ 62 μm after which is the un-deformed matrix (Wu et al. 2002).

The analysis based on the line-broadening principle of XRD peaks leads to the same conclusion of the presence of the crystallite size gradient after SNH treatment (Ortiz et al. 2008). For a C-2000 alloy the crystallite size is found to be 12 nm at the topmost surface layer, which increases gradually to ~ 750 nm at the depth of 500 μm . The crystallite size continues to increase beyond this location until it reaches the depth of 900 μm where the grain size of the undeformed matrix is ~ 50 μm (Ortiz et al. 2008). In addition to the grain/crystallite size gradient, it should be mentioned that the gradients of the dislocation density and lattice microstrains are also present in the surface region, as revealed by detailed XRD analyses (Ortiz et al. 2008).

Work-Hardening Gradient

The microstructure gradient generated by SNH is expected to result in alternation of the local properties at the surface region. One convenient way to probe the change in the local properties is Vickers microhardness measurements or nano-indentation. Shown in Fig. 3 are Vickers microhardness profiles of C-2000 samples as a function of the SNH processing time. Note that the region near the impacted surface exhibits the highest hardness, corresponding to the largest plastic deformation and thus the highest work hardening at that region. This high hardness derives mainly from work hardening because the nc surface layer is too thin (< 5 and ~ 50 μm for the sample processed for 30 and 180 min, respectively) to be measured with a Vickers indenter. The expected hardness gradient is unambiguously present for all the processing times evaluated. However, the most interesting phenomenon to be noted in Fig. 3 is the surface hardening saturation, that is, no additional hardening is obtained beyond 30-min processing. Such a phenomenon has also been observed with an Al-5052 alloy (not shown here). The surface hardening saturation has important



Surface Nanocrystallization and Hardening (SNH), Fig. 3 Vickers hardness profiles of C-2000 samples treated with the SNH process for different periods of time. The hardness profile of an untreated sample is also included for comparison

implications in process optimization because prolonged processing could result in surface contamination of the workpiece through material transfer between the workpiece and impacting balls and thus degrade mechanical properties, particularly the fatigue resistance.

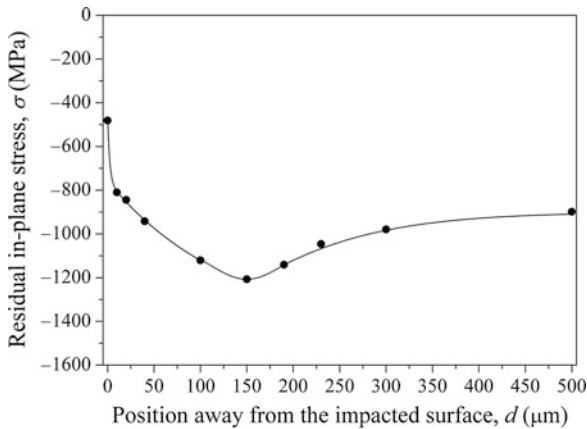
The hardness of the nc surface layer can be a simple extension of or substantially higher than the hardness of the adjacent work-hardened layer, depending on the material under the consideration. With the aid of nano-indentation, it has been shown that for the C-2000 alloy with a FCC crystal structure, the hardness of the nc surface layer is a simple extension of the hardness of the work-hardened layer (Shaw et al. 2008). In sharp contrast, for an eutectoid carbon steel with a pearlite microstructure, the hardness of the nc surface layer is almost twice as high as the value of the adjacent work-hardened region (Umemoto et al. 2004). It is not clear at this stage what are the key factors that determine the hardness value of the nc surface layer relative to that of the adjacent work-hardened region. Additional work is needed to clarify this issue.

Residual Stresses

Plastic deformation on the top layer of the impacted surface not only causes the local property alteration but

also results in the formation of residual compressive stresses. Plastic deformation causes stretching of the top layer of the workpiece. Upon unloading, the elastically stressed subsurface layers tend to recover their original dimensions, but the continuity of the material in both zones, elastic and plastic, does not allow this to occur. As a result, a residual compressive stress field followed by a tensile field is formed in the impacted workpiece. Such phenomena have been well established for the workpiece processed with SP (Al-Obaid 1995; Meguid et al. 1999). The same principles are expected to be operational for SNH-processed components as well.

Figure 4 shows the macroscopic residual stress profile of a C-2000 alloy after SNH treatment for 30 min, determined using XRD based on the peak shifting with respect to the distortion-free reference condition (Ortiz et al. 2008). Note that the maximum residual stress does not appear at the impacted surface, but at the subsurface ($\sim 150 \mu\text{m}$ away from the impacted surface). This residual stress profile is similar to those generated via SP, that is, the maximum residual stress is typically present at the subsurface (Al-Obaid 1995; Meguid et al. 1999). The similarity of the residual stress profiles generated via SNH and SP is not a surprise because both methods entail repeated impacts of the workpiece surface by balls and



Surface Nanocrystallization and Hardening (SNH), Fig. 4

The macroscopic residual stresses on the plane parallel to the impacted surface of a C-2000 alloy after SNH processing for 30 min using WC/Co balls of 7.9 mm in diameter as a function of the position measured from the impacted surface, determined based on the X-ray diffraction combined with material removal layer by layer

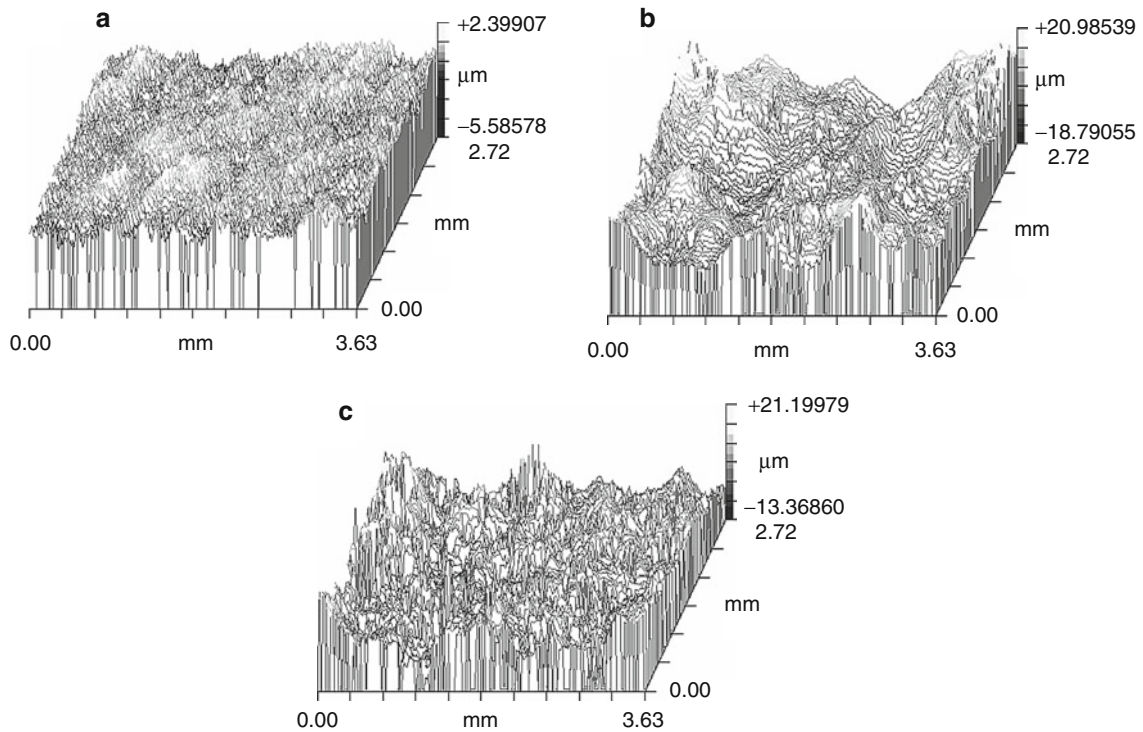
shots. The differences between them are the kinetic energies and sizes of the balls and shots used. These differences, however, do cause differences in residual stresses. As predicted from finite element modeling (Dai and Shaw 2007), the surface layer with residual compressive stresses is much thicker in SNH-processed components than that in the counterparts processed via SP. Furthermore, the maximum residual compressive stress for SNH-processed components is larger than that produced via SP. These larger residual compressive stresses and thicker compressive surface layer are expected to offer better fatigue resistance for SNH-processed components than for SP-processed counterparts.

Surface Roughness and Contamination

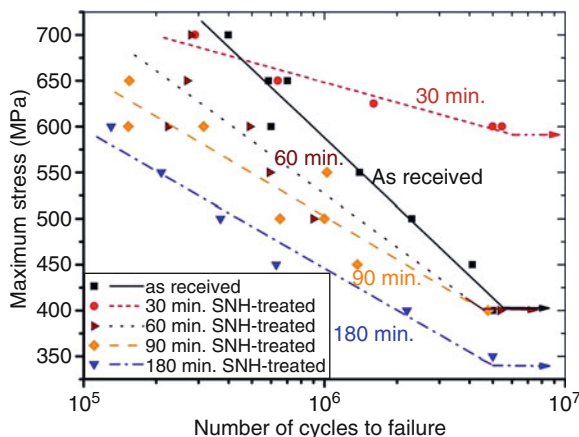
Surface roughness typically increases after SNH processing because of the use of relatively large balls. Detailed studies (Dai et al. 2004a) indicate that surface roughening in the SNH process can be explained by the indentation process of impacting balls, and the surface roughness evolution can be divided into three stages: the roughness increase stage, the roughness decrease stage, and finally the steady-state stage. These three stages are related to different stages of the surface coverage by indents generated by impacting balls. The steady-state stage corresponds to the impact coverage of the entire surface multiple times, and the

surface roughness at the steady-state increases with the ball size (Dai et al. 2004a). The predicted evolution of surface roughness has indeed been confirmed experimentally. As shown in Fig. 5, the sample SNH-processed for 30 min has the highest surface roughness (with the arithmetic-mean value $R_a = 5.5 \mu\text{m}$), followed by the 180-min processed sample ($R_a = 3.6 \mu\text{m}$), with the untreated sample having the lowest surface roughness ($R_a = 0.4 \mu\text{m}$). Thus, surface roughness increases from the untreated sample to the 30-min processed sample, corresponding to the roughness increase stage. In contrast, surface roughness decreases from the 30-min processed sample to the 180-min processed sample, corresponding to the roughness decrease stage. The decrease in surface roughness in this stage is due to the fact that a long time of repeated bombardments by balls can reduce part of the peak height generated by the earlier impacts of balls (Dai et al. 2004a). As a result, R_a increases initially as SNH proceeds, but decreases after long SNH-processing time such as 180 min, as observed in this study.

If cold adhesion between balls and the workpiece takes place during SNH processing, the subsequent separation of balls from the impacted surface will result in surface contamination and may lead to continued increases in surface roughness beyond the steady-state stage as well as introduction of micro-defects. This will inevitably degrade the properties and performance of the impacted workpiece. Studies have indeed revealed that optimization in the SNH time is necessary; otherwise, the fatigue resistance of the impacted workpiece decreases rather than increases. In fact, surface contamination and introduction of micro-defects can have such dramatic effects that the fatigue resistance decreases even though surface roughness decreases with increasing the SNH processing time. Figure 6 shows such a situation. For the 30-min processed C-2000 samples, the fatigue endurance limit has improved from 400 to 600 MPa with respect to the as-received samples without the SNH treatment. However, for the 180-min processed samples, although having lower surface roughness than the 30-min processed samples (Fig. 5), the fatigue endurance limit is, in fact, lower than that of the untreated samples. The 60- and 90-min processed samples exhibit a similar fatigue endurance limit as the untreated samples. These experimental results unequivocally reveal that the beneficial effects from work-hardening, residual compressive stresses, and a nc surface layer can be counterbalanced or even overshadowed by surface contamination and micro-defects if the SNH processing time is not controlled properly. However, this phenomenon is not unique for SNH because “over-processing” has also been observed in SP.



Surface Nanocrystallization and Hardening (SNH), Fig. 5 Three-dimensional surface roughness maps of (a) the annealed C-2000 sample with the arithmetic-mean value $R_a = 0.41 \mu\text{m}$, (b) the sample SNH-processed for 30 min using WC/Co balls of 7.9 mm in diameter with $R_a = 5.50 \mu\text{m}$, and (c) the sample SNH-processed for 180 min with $R_a = 3.64 \mu\text{m}$



Surface Nanocrystallization and Hardening (SNH), Fig. 6 S-N curves for the as-received C-2000 alloy in the annealed condition and SNH-processed counterparts with different processing times, showing that samples with SNH processing for 30 min using WC/Co balls of 7.9 mm in diameter have the highest fatigue endurance limit

Key Applications

Improvements in Friction and Wear Properties

The formation of a nc surface layer and the modification of surface properties induced by SNH can have dramatic impacts on mechanical properties of the workpiece. Friction properties and wear resistance are among those expected to improve because both friction and wear are surface-contact dominated processes. A recent study (Wang et al. 2003) has indeed confirmed such an expectation. It is shown that the friction coefficient of a low-carbon steel is reduced by $\sim 50\%$ after the surface treatment using a SMAT device. The abrasive wear resistance of the same low-carbon steel against a diamond stylus has been improved as well by the SMAT treatment. The improvements in both friction and wear properties have been attributed to the formation of a nc surface layer of $10 \mu\text{m}$ thick induced by SMAT (Wang et al. 2003). It is proposed that the increased hardness associated with the nc surface layer has resulted in shallower penetration of

the diamond stylus into the low carbon steel, and therefore the lower friction coefficient as well as the less wear volume loss due to less plowing and micro-cutting (Wang et al. 2003).

A more complicated situation is found recently with high-carbon steels that are less ductile than low carbon steels (Zhou et al. 2008). It is shown that after the SMAT treatment the high-carbon steel exhibits a similar wear resistance against WC/Co balls as the counterpart without the SMAT treatment, in spite of the formation of a nc surface layer of $\sim 60 \mu\text{m}$ thick in the SMAT-treated steel. However, the wear resistance is markedly enhanced after thermal annealing of the SMAT-treated steel at different temperatures to induce grain growth. Annealing at 650°C results in growth of ferrite grains from 8 to 32 nm, and is particularly effective in enhancing the wear resistance (Zhou et al. 2008). Annealing at temperatures higher than 650°C leads to significant grain growth and the improved wear resistance starts to decrease. These interesting phenomena have been explained in terms of the modified empirical Archard equation (Zhou et al. 2008)

$$W = K \frac{P}{H} \quad (2)$$

where W is the wear rate with an applied pressure P , H is the material's hardness, and K is a pre-factor relative to the material's ductility. The high carbon steel at the SMAT-treated condition has a high hardness with low ductility. As a result, micro-cracking is relatively easy, leading to a high K value and the increased material removal during abrasive wear. After proper annealing at 650°C , the surface of the steel has a lower hardness than the steel at the SMAT-treated condition, but with better plasticity. As such, the steel annealed at 650°C exhibits the best wear resistance because of the proper combination of sufficient hardness and moderate plasticity. In contrast, the steel annealed at temperatures higher than 650°C is too soft to resist the abrasive wear and thus has the decreased wear resistance (Zhou et al. 2008).

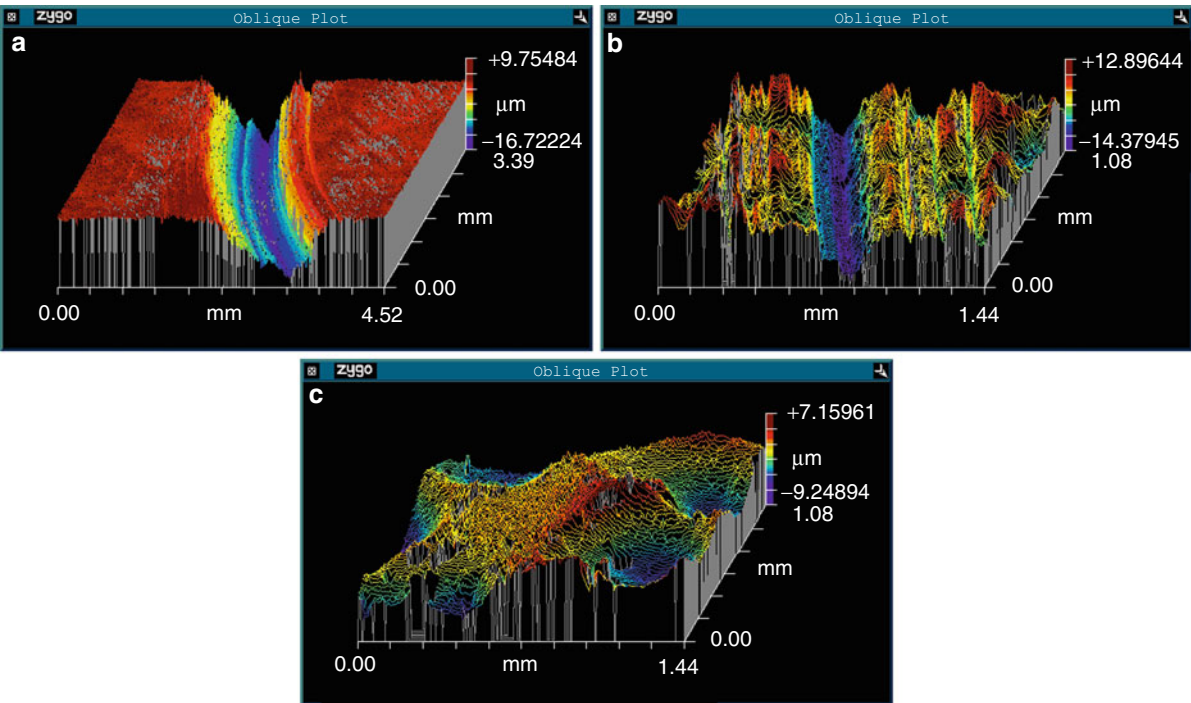
Friction and wear properties of the nickel-based C-2000 alloy with the SNH treatment have also been examined. Shown in Fig. 7 are three-dimensional wear tracks of the annealed, shot-peened, and SNH-processed C-2000 alloy. It is clear that the annealed C-2000 alloy has the deepest groove and thus the worst wear resistance, while the SNH-processed C-2000 alloy has the best wear resistance because it exhibits a negligible worn track. Quantitative analysis of the worn track reveals that the wear rate of the annealed C-2000 alloy against Si_3N_4 balls is $1 \times 10^{-5} \text{ mm}^3/\text{N.m}$. The shot peening treatment

has reduced the wear rate by one order of magnitude to $1 \times 10^{-6} \text{ mm}^3/\text{N.m}$, whereas the wear rate of the SNH-processed sample is negligible because the wear volume is too small to be measurable with accuracy under the testing condition investigated. As shown in Fig. 8, the SNH-processed C-2000 alloy also exhibits the lowest friction coefficient, followed by the SP-processed sample with the annealed sample having the highest friction coefficient. These results reveal that the SNH-processed C-2000 alloy has the best friction and wear resistance properties, whereas the annealed counterpart has the worst properties, with the SP-processed sample between these two conditions. This ranking can be attributed to the work-hardened surface in the SP-processed sample and the formation of a nc surface layer in the SNH-processed counterpart.

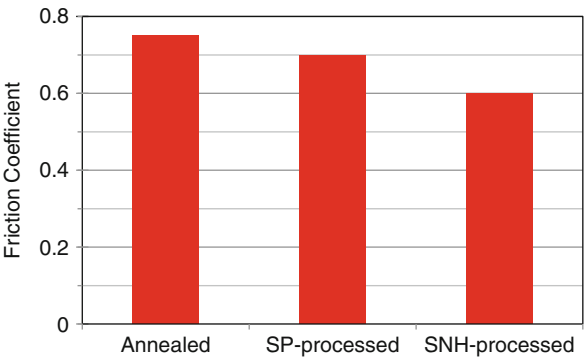
It should be mentioned that improved wear resistance has been reported previously by many studies on shot peened samples, and the improvement has been related to one or all of the following mechanisms: (a) the presence of residual compressive stresses, (b) the increased hardness, (c) the decreased coefficient of friction, and (d) the altered surface roughness. The function of compressive stresses is to close up micro-cracks at the surface induced by wear and fretting and prevent them from propagation. The increased hardness decreases the penetration into the workpiece by asperities of the antagonist, and thus increases the abrasive wear resistance, decreases the friction coefficient, and may prevent adhesion with the antagonist. Surface roughness plays dual functions; it could reduce wear and fretting resistance if a rough surface acts as potential stress raisers. However, a high degree of surface finishing can increase the friction coefficient and thus accentuate wear and fretting damage. The dual functions of surface roughness underscore the importance of process optimization because not all shot-peened materials exhibit improvements. Currently, detailed studies of the contribution from each factor listed above in conjunction with the presence of a nc surface layer have not been carried out yet for SNH-processed materials. Systematic studies in this area are anticipated in the near future.

Improvements in Other Mechanical Properties

The improvements in mechanical properties other than friction and wear resistance via the SNH treatment have also been observed in many materials. In particular, the fatigue resistance has been enhanced substantially (Villegas et al. 2005). As shown in Fig. 6, SNH processing of a C-2000 alloy for 30 min can lead to a 50% improvement in the fatigue endurance limit. It is well known that



Surface Nanocrystallization and Hardening (SNH), Fig. 7 Three-dimensional wear tracks of the C-2000 alloy with a Si_3N_4 ball as the antagonist: (a) the annealed, (b) the shot-peened, and (c) the SNH-processed sample. The wear test was conducted in a pin-on-disk configuration under ambient laboratory conditions (20°C and 45% relative humidity) and a load of 4.9 N without lubrication. Note that the deepest groove is created in the annealed sample, while the surface of the SNH-processed sample remains almost the same as that before the wear test



Surface Nanocrystallization and Hardening (SNH), Fig. 8 A comparison of the friction coefficients of the annealed, shot-peened, and SNH-processed C-2000 samples. The friction coefficients were determined using the conditions described in Fig. 7

shot peening also results in improvements in the fatigue endurance, and it has been used widely because of its flexibility in treating components of simple and complex

geometries, serving as the gold standard for the last 50 years for fatigue resistance applications. Currently, 75% of components in airplane engines are subjected to shot peening. The improvement in the case of shot peening is generally attributed to the presence of residual compressive stresses at the surface region. However, with the aid of finite element modeling, it is found that the nc surface layer and the work-hardened surface region in SNH-processed components play more significant role in enhancing the fatigue endurance limit than residual compressive stresses. Because of the co-presence of significant residual compressive stresses, the nc surface layer, and a thick work-hardened region, it is expected that SNH can offer larger improvements than SP can. Studies in this area are currently underway to explore the full potential of SNH in enhancing the fatigue resistance of various materials.

It should be noted that although SNH can improve the fatigue resistance of materials, “over-processing” can result in degradation of the fatigue resistance rather than improvements, as shown in Fig. 6. The phenomenon of

“over-processing” is due to surface contamination and micro-defects introduced during the SNH treatment. Surface roughness does not seem to play a key role in the degraded fatigue resistance shown in Fig. 6 because the C-2000 alloy SNH-processed for 180 min has a lower surface roughness than the counterpart SNH-processed for 30 min. Therefore, if surface contamination and the formation of micro-defects can be avoided through the judicial selection of appropriate balls, further enhancements in the fatigue endurance limit beyond what is shown in Fig. 6 are possible.

Since the SNH process only alters surface microstructure and thus surface properties, it is expected that SNH cannot change the tensile strength of bulk materials significantly. However, when bulk materials are in a plate form (or a rod geometry), a 35% improvement in the tensile yield strength of mild steels with minimum degradation in ductility and toughness has been demonstrated. A recent study (Tian et al. 2008) indicates that about 100% improvements in the tensile yield strength of the nickel C-2000 alloy can be achieved by SNH. Furthermore, it is noted that the ultimate tensile strength of the nickel C-2000 alloy is also increased by SNH, suggesting that the improvement in tensile strength is not due to work hardening alone, but also due to the contribution from the nc surface layer. However, accompanied with the increased tensile strength, the tensile ductility is decreased. As shown in Fig. 9, the longer the SNH processing time, the

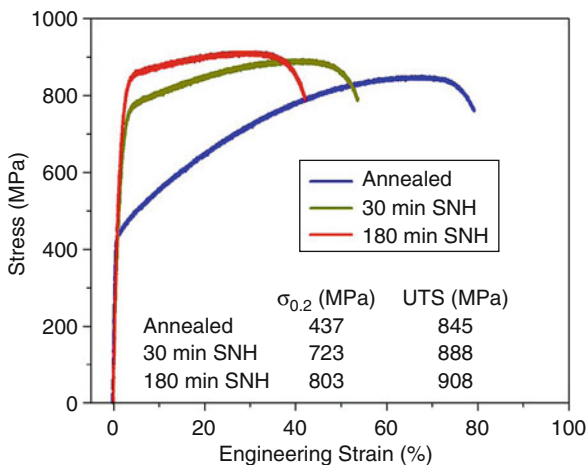
more reduction in the ductility. In spite of the reduction, the remaining ductility after reduction ($\sim 40\%$) is still more than sufficient for engineering applications. Detailed finite element modeling coupled with the analyses of the stress-strain curves and fracture surface (Tian et al. 2008) reveal that the reduced ductility is mainly related to the decreased strain-hardening coefficient, which, in turn, is caused by the presence of the nc surface layer and the work-hardened surface region created via SNH. The residual compressive stresses introduced by SNH have little influence on the tensile ductility. Instead, residual compressive stresses extend the apparent elastic strain of SNH-processed samples. The surface roughness of SNH-processed samples does not have much influence on the ductility either. However, surface roughness may play a role in changing the cup-and-cone fracture for the annealed sample to the shear fracture for SNH-processed samples (Tian et al. 2008).

Acknowledgments

The author would like to thank his former students, Dr. Juan C. Villegas, Dr. Kun Dai, and Mr. Misael Manjarres for their significant contributions to the work present here. The collaboration with and contributions from Prof. Peter K. Liaw and Mr. Jiawan Tian at the University of Tennessee, Prof. Angel L. Ortiz at the Universidad de Extremadura, Spain, and Dr. Dwaine L. Klarstrom at Haynes International are greatly appreciated. The financial support from the National Science Foundation through Grant No. DMR-0207729 is acknowledged.

References

- Y.F. Al-Obaid, Shot peening mechanics: experimental and theoretical analysis. *Mech. Mater.* **19**, 251 (1995)
- K. Dai, L. Shaw, Comparison between shot peening and surface nanocrystallization and hardening processes. *Mater. Sci. Eng. A* **463**, 46 (2007)
- K. Dai, J.C. Villegas, Z. Stone, L. Shaw, Finite element modeling of the surface roughness of 5052 Al alloy subjected to a surface severe plastic deformation process. *Acta Mater.* **52**, 5771 (2004a)
- K. Dai, J.C. Villegas, L. Shaw, An analytical model of the surface roughness of an aluminum alloy treated with a surface nanocrystallization and hardening process. *Scr. Mater.* **52**, 259 (2004b)
- S.A. Meguid, G. Shagal, J.C. Strannart, J. Daly, Three-dimensional dynamic finite element analysis of shot-peening induced residual stress. *Finite Elem. Anal. Design* **31**, 179 (1999)
- A.L. Ortiz, J.W. Tian, J.C. Villegas, L. Shaw, P.K. Liaw, Interrogation of the microstructure and residual stress of a nickel-base alloy subjected to surface severe plastic deformation. *Acta Mater.* **56**, 413 (2008)
- M.C. Shaw, G.J. DeSalvo, On the plastic flow beneath a blunt axisymmetric indenter. *Trans. ASME, J. Eng. Ind.* **92**, 480 (1970)
- L. Shaw, J.C. Villegas, J.Y. Huang, S. Chen, Strengthening via deformation twinning in nickel alloys. *Mater. Sci. Eng. A* **480**, 75 (2008)



Surface Nanocrystallization and Hardening (SNH), Fig. 9 Engineering stress-strain curves of C-2000 specimens before and after SNH treatment with different times indicated. The values of the yield stress, $\sigma_{0.2}$, and the ultimate tensile strength, UTS, of various samples are listed for comparison

- N.R. Tao, Z.B. Wang, W.P. Tong, M.L. Sui, J. Lu, K. Lu, An investigation of surface nanocrystallization mechanism in Fe induced by surface mechanical attrition treatment. *Acta Mater.* **50**, 4603 (2002)
- J.W. Tian, L. Shaw, P.K. Liaw, K. Dai, On the ductility of a surface severely plastically deformed nickel alloy. *Mater. Sci. Eng.* **498**, 216 (2008)
- M. Umamoto, K. Todaka, K. Tsuchiya, Formation of nanocrystalline structure in carbon steels by ball drop and particle impact techniques. *Mater. Sci. Eng. A* **375–377**, 899 (2004)
- J.C. Villegas, L. Shaw, K. Dai, W. Yuan, J.W. Tian, P. Liaw, D.L. Klarstrom, Enhanced fatigue resistance of a nickel-based Hastelloy induced by a surface nanocrystallization and hardening process. *Phil. Mag. Lett.* **85**, 427 (2005)
- Z.B. Wang, N.R. Tao, S. Li, W. Wang, G. Liu, J. Lu, K. Lu, Effect of surface nanocrystallization on friction and wear properties in low carbon steel. *Mater. Sci. Eng. A* **352**, 144 (2003)
- X. Wu, N. Tao, Y. Hong, B. Xu, J. Lu, K. Lu, Microstructure and evolution of mechanically induced ultrafine grain in surface layer of Al-alloy subjected to USSP. *Acta Mater.* **50**, 2075 (2002)
- L. Zhou, G. Liu, Z. Han, K. Lu, Grain size effect on wear resistance of a nanostructured AISI52100 steel. *Scr. Mater.* **58**, 445 (2008)

Surface Physics Concepts

- [Surface Forces, Surface Tension, and Adhesion](#)

Surface Quality

- [Tribological Effects of Machining Carbon Nanotube Composites](#)

Surface Roughness

- [Surface Statistics and Probability Density Function](#)

Surface Roughness of Bearing Components

- [Rolling Element Bearing Surface Finish](#)

Surface Smoothness of Bearing Components

- [Rolling Element Bearing Surface Finish](#)

Surface Statistics and Probability Density Function

HORST BODSCHWINNA¹, JÖRG SEEWIG²

¹Institut of Measurement and Automatic Control, Leibniz University Hannover, Hannover, Germany

²Lehrstuhl für Messtechnik & Sensorik, Technische Universität Kaiserslautern, Kaiserslautern, Germany

Synonyms

[Mean roughness of a surface \(\$R_a\$ \)](#); [Roughness from surface statistics](#); [Surface roughness](#); [Vertical and horizontal roughness parameters](#)

Definition

The description of a profile's properties comprises the statistically independent identification of (1) properties of height and (2) horizontal properties.

In industrial praxis, measures of height are preferable, for they are seen in close relation with the tolerance of fit. For surfaces with high quality requirements, additional horizontal and hybrid measurements of roughness are also used.

Scientific Fundamentals

Surface Statistics and 2D Parameters

For characterization of height, either “peak-value-orientated” parameters, like the maximum roughness height R_z , or average values, like the arithmetical mean deviation of the roughness profile R_a and the root mean square deviation of the roughness profile R_q , are applied.

Accordingly, the following facts have to be taken into account when selecting the measuring section:

- The value of “peak-value-orientated” parameters decreases monotonously with the measuring length l . Thus, this characteristic is defined in ISO 4287 (ISO 4287 Geometrical Product Specifications (GPS) 1997) within the sampling length, so the measuring section cannot be chosen independently.
- Average values, on the other hand, converge on the real value with increasing measuring length and become statistically safer by ensemble averaging.

Due to its high statistical safeness, the roughness measure R_a is used in western countries in the field of quality control. Presently, the following parameters are generally used in industrial applications:

- Rz , maximum roughness height
- Rt , total height of roughness
- Ra , arithmetical mean deviation
- Rq , root mean square deviation
- R , average value of height of MOTIF

In Germany, the most commonly used are the ISO 4287 (ISO 4287 Geometrical Product Specifications (GPS) 1997) standardized parameters Rz and Ra , for which standardized measurement conditions are also available. The MOTIF parameter R , which was developed by the French automotive industry, was standardized in ISO 12085 (ISO 12085 Geometrical Product Specifications (GPS) 1997) and uses its own filtering technique. The root mean square deviation Rq of the profile is calculated from the roughness profile coordinates $r(x)$.

$$Rq = \sqrt{\frac{1}{l} \int_0^l r(x)^2 \cdot dx} \quad (1)$$

Thus, Rq equals the effective value of the profile's ordinate and the deviation σ of the Gaussian distribution. But the roughness parameter Rq is no longer in use. Ra has slightly smaller values compared with Rq because the contribution of the profile's ordinates is taken into account:

$$Ra = \frac{1}{l} \int_0^l |r(x)| dx \quad (2)$$

Horizontal and hybrid properties are valuable additional information, especially when the properties of the function are co-determined by the structural properties of the roughness. For instance, the deep drawing properties of rolled thin sheets are greatly dependent on the horizontal properties. Therefore, high-spot-count methods (also known as peak-count methods) are applied. However, the results of devices for different manufacturers cannot currently be compared because, in addition to the distance between the counting level and the center line, the vertical resolution of the instrument and a threshold for peak detection at the counting level are applied have to be established consistently.

According to Whitehouse (1994), for the tribological behavior of a surface, two hybrid characteristics of the profile are of great importance: slopes and bends. These can be calculated from the derivatives of the profile.

Corresponding statistical parameters are the root mean square slope of the assessed profile $R\Delta q$

$$R\Delta q = \sqrt{\frac{1}{l} \int_0^l \left| \frac{\partial r(x)}{\partial x} \right|^2 \cdot dx} \quad (3)$$

and the squared mean value of the bends ρq

$$\rho q \approx \frac{1}{l} \int_0^l \left| \frac{\partial^2 r(x)}{\partial x^2} \right| dx \quad \text{with} \quad \frac{\partial z}{\partial x} \ll 1 \quad (4)$$

Whitehouse (1994) points out that a distinction between slopes and bends in peak and groove regions is necessary for functional reasons. This idea of the different weighting of peak and groove regions is realized in the internationally standardized three-straight-line model of Abbott's material fraction curve, which separates roughness height into regions of functionally different effectiveness.

Amplitude-Density Function

Figure 1 gives the schema of the determination of the amplitude density with the help of a probability analysis of the profile amplitudes. In the case of a continuous signal and for a limit $\Delta z \rightarrow 0$, the histogram passes into the continuous amplitude-density function. The integrated amplitude density over the profile range has the value "1," i.e., the area under the curve is "1."

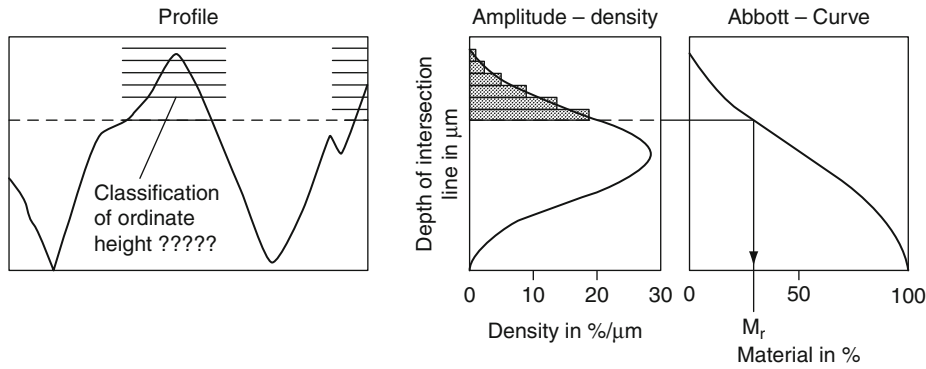
$$\int_{-\infty}^{\infty} p(z) \cdot dz = 1 \quad (5)$$

The variety of manufacturing methods leads to various shapes of the amplitude-density function, which describes the character of the surface. On the right-hand side in Fig. 1, the Abbott curve is plotted. It was named after Abbott and Firestone (1933) and evolves from the amplitude-density function by integration, beginning at the peaks in the profile. For its simple interpretation, this function is used in characterization of surfaces.

The central moments of n th order act as parameters for the amplitude-density function.

$$\begin{aligned} \mu_n &= E[(z - E[z])^n] = \int_{-\infty}^{+\infty} (z - E[z])^n \cdot p(z) \cdot dz, \\ E[z] &= \int_{-\infty}^{+\infty} z \cdot p(z) \cdot dz \end{aligned} \quad (6)$$

Because a high-pass filter is used for the roughness measurement, the arithmetical mean value of the amplitudes is $E[z] = 0$. The 2nd central moment represents the variance σ^2 of the profile amplitudes. The skewness Rsk



Surface Statistics and Probability Density Function, Fig. 1 Correlation between profile, amplitude-density function and Abbott curve

characterizes the asymmetry of the density distribution and is defined by the central moments of 2nd and 3rd order as stated:

$$Rsk = \frac{\mu_3}{\mu_2^{3/2}} \quad (7)$$

With a symmetrical density function, the skewness equals zero. Asymmetrical density functions have the value $Rsk \neq 0$. The kurtosis or steepness Rku marks the profile as oblate or afflicted with outliers.

$$Rku = \frac{\mu_4}{\mu_2^2} \quad (8)$$

In doing so, the steepness is judged in comparison to the steepness $Rku = 3$ of the Gaussian distribution. $Rku < 3$ points to flatter and $Rku > 3$ to steeper density distributions (Fig. 2).

Whitehouse (1994) notes the difficulties that result from the statistical dependency of these parameters. The beta function, which was introduced by Whitehouse (1978) for approximation of amplitude density functions, has two parameters, a and b , that are statistically independent of each other

$$\beta(a, b) = \int_0^1 z^{a-1} \cdot (1-z)^{b-1} \cdot dz \quad (9)$$

Therefore, the amplitude density function is normalized to range 1. The parameters a and b are determined by height parameters from measured profiles:

$$a = \frac{Rv(Rv \cdot Rp - Rq^2)}{Rt \cdot Rq^2}, \quad b = \frac{Rp(Rv \cdot Rp - Rq^2)}{Rt \cdot Rq^2} \quad (10)$$

Here,

Rt = total height of roughness (range of the distribution),

Rp = maximum profile peak height,

Rv = maximum profile valley depth.

Prerequisite for the informational value of the beta function is the statistically safe determination of these limit-orientated parameters.

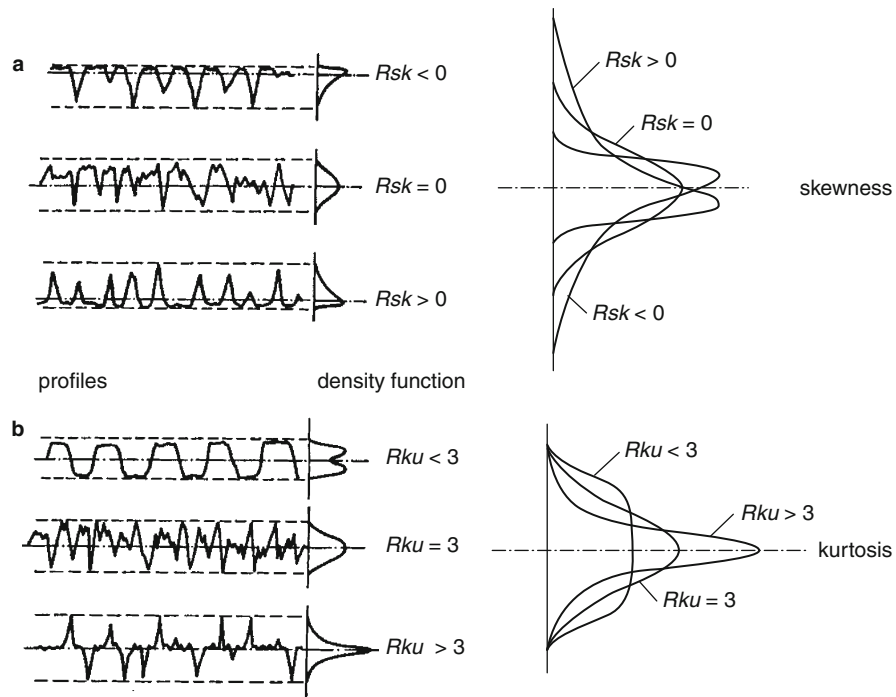
Parameters for Characterization of the Abbott Material Fraction Curve according to ISO 13565, Part 2

This model for the description of technical surfaces is based on the following fundamental perceptions:

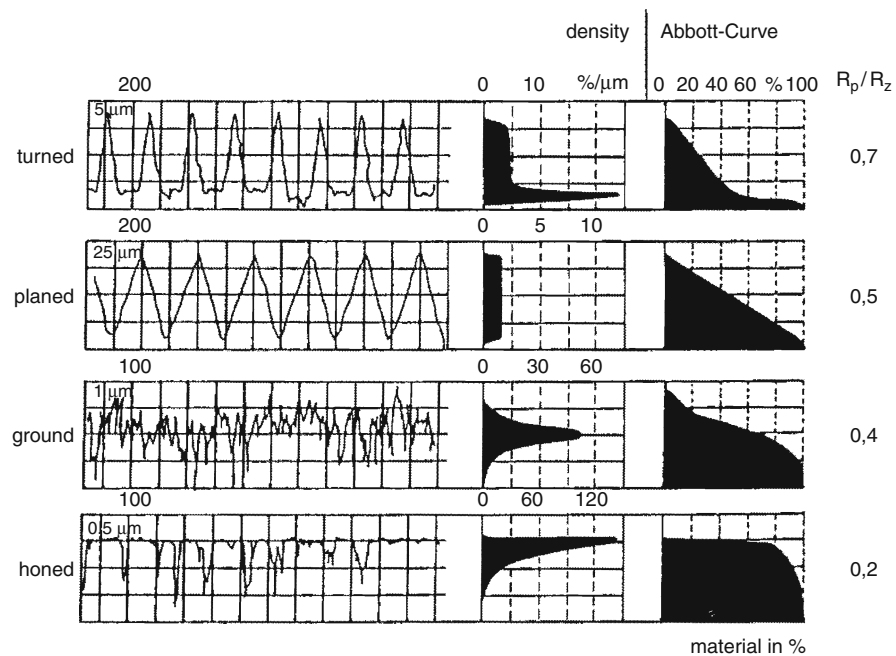
- The functional properties of a surface are not only determined by the depth of roughness but, to a great extent, also by the character of the micro-geometry caused by production.
- Elevations and depressions in roughness have a different influence on function; hence, judging the level of such elements must take the overall structure into account.
- The mechanical strength of technical surfaces depends not only on material properties but also on material distribution across the depth of the roughness structure.

Figure 3 reflects the surface profiles of different manufacturing methods that show different material distributions accordingly. Since the publication by Abbott and Firestone (1933) in 1933, these characteristic functions are also called Abbott-Firestone curves, or simply Abbott curves.

In the upper section of Fig. 3, two profiles with an “open” profile character are shown, and in the lower section “closed” and thus more functional and



Surface Statistics and Probability Density Function, Fig. 2 Characterization of the shape of the amplitude density curve by (a) skewness and (b) kurtosis



Surface Statistics and Probability Density Function, Fig. 3 Shapes of profiles, amplitude density curves and Abbott curves for different manufacturing methods

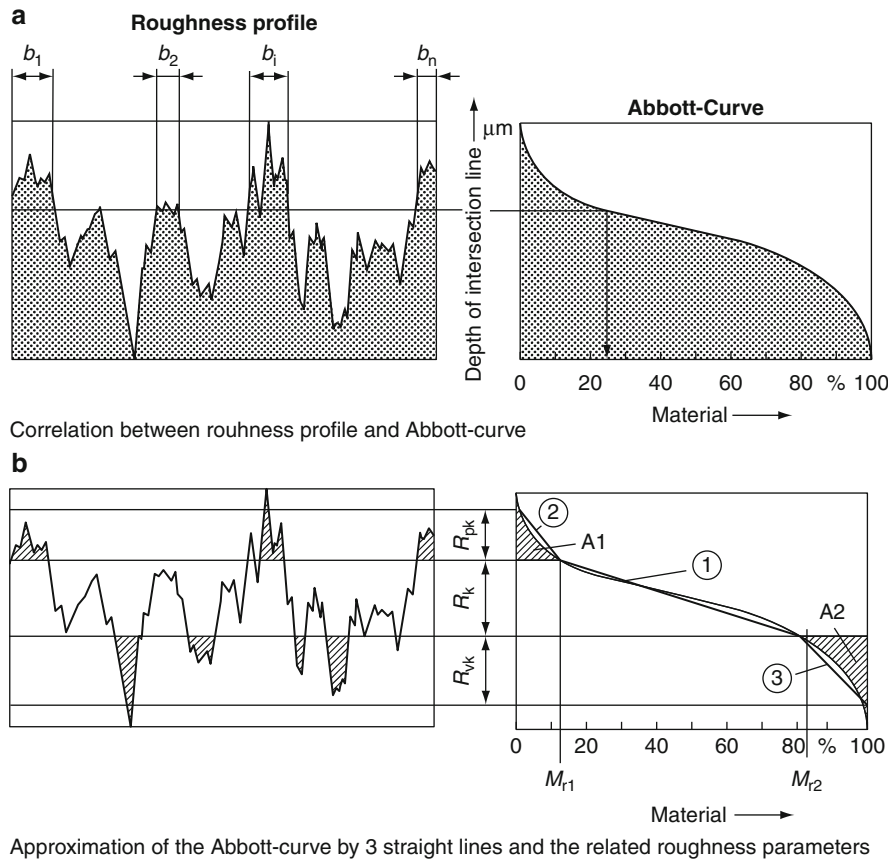
advantageous profiles are displayed. Regarding optimal functional properties of surfaces subject to high mechanical stress, asymmetric profiles with a plateau-like profile ratio are the goal, as they are expressed in Fig. 3 using the example of a honed surface. These functionally relevant properties lead to characteristic shapes in amplitude density curves and Abbott curves.

In as much as the Abbott curve results from the amplitude density by integrating over the depth of roughness, both curves have the same informational value. Evaluating shape and depth of the Abbott curve offers the advantage of very simple visualization of the surface's properties in relation to the technical performance in use.

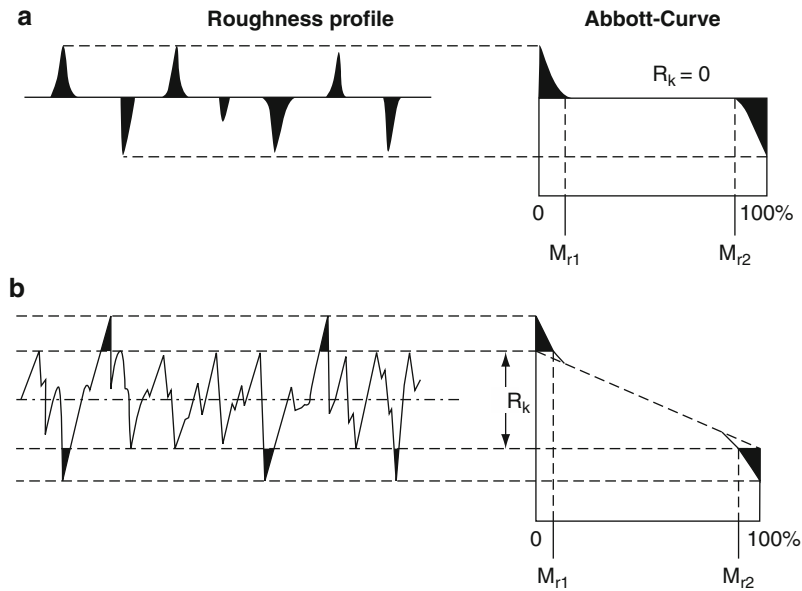
Figure 4 shows the development and meaning of the parameters in the Abbott curve in connection with the roughness profile. In the upper section, the relation between the roughness profile and the Abbott curve, i.e., the share of material in the roughness profile as a function of the position of the line of intersection, is displayed.

The lower section of the Fig. 4 clarifies how a classification of the profile's roughness depth into (a) the profile peak area, (b) the profile core area, and (c) the profile groove area can be achieved by approximation of the Abbott curves by the straight lines 1, 2, and 3. These areas are important to the functional behavior and are therefore described by separate parameters. The main parameters are (a) average height of the protruding peaks above the roughness core profile Rpk , (b) depth of the roughness core profile Rk , and (c) average depth of the profile valleys projecting through the roughness core profile Rvk .

A universal physical definition for classification of the roughness profile in a peak area, a core area, and a groove area is not available. The basic idea of the method defined in ISO 13565 (ISO 13565-2 Geometrical Product Specifications (GPS) 1996) is the classification of the roughness profile into three areas with significantly different material increase. The classification is done by means of the approximation of the Abbott curve by three straight lines.



Surface Statistics and Probability Density Function, Fig. 4 Characterization of surfaces according to ISO 13565, Part 2 (ISO 13565-2 Geometrical Product Specifications (GPS) 1996)



Surface Statistics and Probability Density Function, Fig. 5 Influence of the core area's properties on the determination of the peak and groove areas

The core area distinguishes itself by the highest material increase with advancing cutting depth in the profile. Therefore, the core area has a considerably higher mechanical strength and a higher wear resistance. The Abbott curve has the smallest gradient in this area of highest material increase. In case of a geometrically ideal surface or an geometrically ideal plateau with projecting peaks and grooves, the straight line (1) is horizontal and the core roughness depth R_k equals zero (upper section of Fig. 5).

Only in this case if the plateau has an ideal geometry ($R_k = 0$) a sharp separation of peaks and grooves is applicable. With a high core roughness depth, as seen in the lower section, only a considerably smaller part of the profile is counted as projecting peaks and grooves. This is valid for judging the functional behavior because there are already relevant elevations and depressions in the core area. Only very projecting outliers, and here especially extreme profile peaks, have an additional influence on the functional characteristic of contact surfaces.

The specification of micro-geometrical properties by the designer starts with the specification of the core roughness depth R_k and thus with the material increase in the core area. The smaller the R_k value, the more resistant is the surface in the core area. Additionally, the values Mr_1 and Mr_2 , which are the limits of the core area, are specified. Within these limits R_k exists with the specified properties.

The reduced peak height Rpk marks the proportion of the projecting profile peaks that stick out of the core profile. This parameter describes profile properties used for judging the running-in characteristic, for instance, of lubricated sliding and rolling faces. The goal is, for example, at honed surfaces, a small Rpk value that results in a forestalling of the running-in wear by the manufacturing method.

The reduced groove depth Rvk characterizes the proportion of the profile depths that extend into the material from the core profile. In the case of forged functional surfaces, e.g., for cylinder liners, a high Rvk value is desirable in order to enable absorption of lubricant with an otherwise plateau-like structure.

The parameters Rpk and Rvk are calculated as a side length of a triangle that is equal in area with the peak area $A1$ and the groove area $A2$ (Fig. 5). In Fig. 5, $A1$ matches the filled profile peaks and $A2$ the unfilled grooves. Even though the flank parameters $A1$ and $A2$ bear a direct relation to judging the running-in characteristic and the absorption of lubricant, it seems easier in terms of handling the parameters in practice to convert them to vertical quantities. The parameters Rpk and Rvk facilitate (compared with flank parameters) weighting the peak and groove areas in relation to core roughness depth.

Thus, the fundamental parameter for rating functional behaviour is always the core roughness depth R_k , i.e., the characteristic of the core area. The dimension of the core

roughness depth and the proportionally existent peak and groove regions are therefore the crucial parameters in rating the functional properties of the surface. The core roughness depth R_k is determined with the help of a best-fit line comprising 40% of the profile ordinates in the flattest progression of the Abbott curve (Fig. 6).

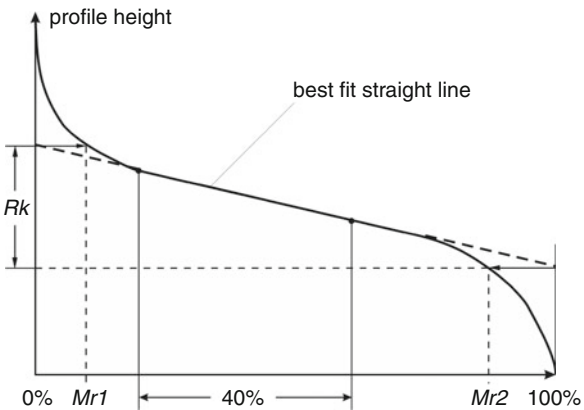
This descriptive model is all the more informative the more distinct the plateau characteristics of a surface are. The material fraction difference of 40% was established

empirically after evaluation of many industrially produced functional surfaces. It is grounded on the requirement that mechanically highly stressed surfaces should feature a definite fraction of at least 40% in the core area. A core roughness depth of $R_k = 2\text{ }\mu\text{m}$ can be interpreted as an increase in material by 10% when advancing $0.2\text{ }\mu\text{m}$ into the “core” profile.

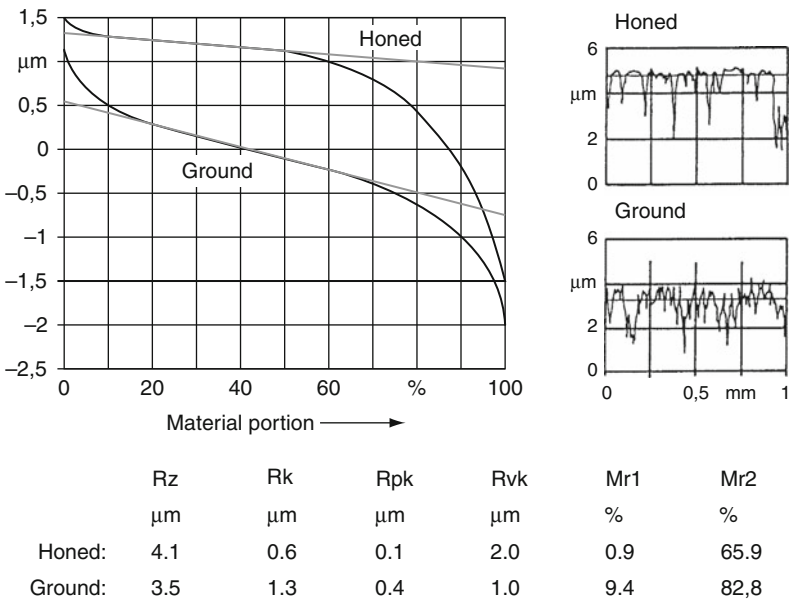
Key Applications

Characterization of Ground and Honed Surfaces according to ISO 13565, Part 2

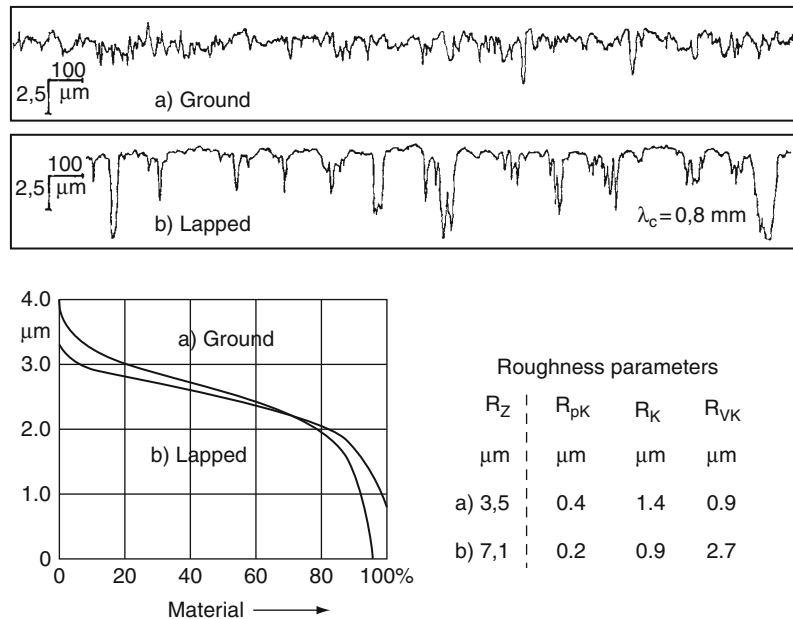
Figure 7 illustrates the meaningfulness of these parameters by comparison of a ground and a honed surface. Both have the nearly the same R_z value but very different profile characters. The required best-fit lines for determination of the R_k value is plotted in dashes to each Abbott curve. The Abbott curve of the honed surface shows a flat progression over a long range, which leads to a core roughness depth R_k of only $0.6\text{ }\mu\text{m}$. Comparing this to the average surface roughness $R_z = 4.1\text{ }\mu\text{m}$, the plateau-like character of the surface is expressed. Furthermore, the very small value of $R_{pk} = 0.1\text{ }\mu\text{m}$ proves the directed removal of the profile peaks by honing. In contrast, the ground surface has a significantly higher core roughness depth and with $R_{pk} = 0.4\text{ }\mu\text{m}$ a noticeable higher peak area. These two properties explain the unsatisfactory functional behaviour



Surface Statistics and Probability Density Function, Fig. 6 Identification of the core area and the parameters R_k , Mr_1 , and Mr_2



Surface Statistics and Probability Density Function, Fig. 7 Abbott curves and parameters of a ground and a honed surface with nearly the same roughness depth



Surface Statistics and Probability Density Function, Fig. 8 Measurements of roughness depth on bath nitrided and lapped camshafts made of cast iron (Volkswagen AG, Salzgitter)

of the ground surface despite the marginally lower R_z value of 3.5 μm .

Typical for the honed surface is the reduced groove depth $Rvk = 2 \mu\text{m}$ compared with the half-as-high value at the ground surface. The Rvk value is vital for the absorption of lubricant.

Bearing Seats of Camshafts Made of Cast Iron

It is possible to characterize the quality of the surface finishing even for materials like cast iron or porous sinter materials. This is achieved by differentiation of the complete roughness depth into areas of different roughness depth.

Figure 8 returns roughness measurements on bearing seats of camshafts of pearlitic cast iron in different stages of manufacturing. In pre-ground condition, the geometrical properties of the roughness can be described properly by the R_z parameter. Following the bath nitriding and subsequent belt lapping, graphite components have dissolved from the surface so that an evaluation of the surface finishing achieved by lapping is pointless.

A comparison of the Abbott curves, on the other hand, allows for clear identification of the improvement in surface quality by a far flatter progression of the Abbott curve in the core area and a much smaller peak area. For quality control, the sum of values of Rpk and Rk is used, which decreases significantly due to belt lapping. The dissolving

of graphite components yields a higher groove and pore area that comes, in the case of a sufficiently high $Mr2$ value, with advantageous functional properties for a sufficient absorption of lubricant.

The parameters of the Abbott curve only include vertical properties of the surface and have to be extended by adequate parameters from the wavelength spectrum or the autocorrelation function if an influence on the function by horizontal properties is expected.

Cross-References

- [Filtration of Surface Measurement Data](#)
- [Surface Characterization and Description](#)
- [Surface Roughness](#)
- [Surface Synthesis Based on Surface Statistics](#)
- [Surface Variation in Tribological Processes](#)

References

- E.J. Abbott, F.A. Firestone, Specifying surface quality. *Mech. Eng.* 55, 569–572 (1933)
- ISO 12085 Geometrical Product Specifications (GPS), Surface texture: profile method – motif parameters (1997)
- ISO 13565-2 Geometrical Product Specifications (GPS), Surface texture: profile method – surfaces having stratified functional properties – part 2: height characterization using the linear material ratio curve (1996)
- ISO 4287 Geometrical Product Specifications (GPS), Surface texture: profile method – terms, definitions and surface texture parameters (1997)

- D.J. Whitehouse, Beta functions for surface typology. *Ann. CIRP* 27, 491–497 (1978)
- D.J. Whitehouse, *Handbook of Surface Metrology* (Institute of Physics, Bristol, 1994). ISBN 0-7503-0039-6, 988

Surface Synthesis Based on Surface Statistics

FEODOR M. BORODICH¹, DAVIDE BIANCHI²

¹School of Engineering, Cardiff University, Cardiff, Wales, UK

²Austrian Center of Competence for Tribology AC²T, Wiener Neustadt, Austria

Synonyms

Modeling of surface roughness

Definition

Surface synthesis is the process of mimicking a natural surface in order to create a synthetic surface whose main characteristics of topography are the same from a statistical point of view as the characteristics of the original surface.

Scientific Fundamentals

Historically, problems of tribology were studied initially for bodies of classical shapes (e.g., spheres, cones, and flat-ended punches) having ideal smooth surfaces. However, it was soon realized that deviations of contact surface from ideal shapes and the effect of surface features such as bumps, waviness, and roughness can have a great influence on the results of calculations. Currently modern experimental techniques (e.g., techniques based on atomic-force microscopy) allow researchers to describe the surface topography up to atomic scale resolution. However, even if it were possible to compute a tribological problem for a real rough surface, this would be of little use, as it would not even allow the prediction of the problem behavior for a second rough surface or even for the same surface but with roughness profile measured at a slightly different place (Lubrecht and Venner 1999). Indeed, for such a prediction, one has to understand which parameters of the rough surfaces are the governing parameters for the process under consideration and be able to predict the changes in the process behavior when the parameters are varied. The use of synthetic surfaces has great potential value to researchers because

it allows modeling of various tribological phenomena such as lubrication, wear, etc. Thus, one has to be able to model the main features of rough surfaces and to understand the behavior of the features at different scales of the real roughness in order to improve the predictive capacities of engineers concerning tribology of real surfaces.

Today, more than 30 parameters and functions are used in order to characterize the complex structures of rough surfaces. In particular, these parameters include (i) height (amplitude) parameters, e.g., the maximum height of the profile; (ii) horizontal parameters, e.g., the number of intersections of the profile with the mean line; (iii) parameters related to the shape of protuberances, e.g., the root mean square (rms) slope and the rms curvature of the profile; and (iv) parameters associated with spatial extend and amplitude of the roughness, like the high spot count (Nowicki 1985). The most common statistical height parameters for a profile function $z(x)$ within an interval $[-L, L]$ are the arithmetical mean deviation of the profile R_a , the maximum height of the profile R_{max} , and the rms roughness height R_q :

$$R_a = \frac{1}{2L} \int_{-L}^L |z(x)| dx, \quad R_{max} = \max_{x \in [-L, L]} z(x),$$

$$R_q = \left[\frac{1}{2L} \int_{-L}^L [z(x)]^2 dx \right]^{1/2}.$$

Whitehouse and Archard (1970) suggested studying surface roughness using the auto-correlation function $R(\delta)$ and its Fourier transform, the power spectral density $G(\omega)$:

$$R(\delta) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T [z(x+\delta) - \bar{z}] [z(x) - \bar{z}] dx$$

$$= r\langle [z(x+\delta) - \bar{z}] [z(x) - \bar{z}] \rangle,$$

$$G(\omega) = \frac{2}{\pi} \int_0^\infty R(\delta) \cos \omega \delta d\delta \quad \text{and}$$

$$\bar{z} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T z(x) dx$$

where \bar{z} is the average value (the mean line) of the profile function $z(x)$. In fact, $R(\delta)$ and $G(\omega)$ are the main tools for studying statistical models of rough surfaces. The moments m_n of the spectral density $G(\omega)$ provide a useful description of the surface roughness

$$m_n = \int_{\omega_0}^\infty \omega^n G(\omega) d\omega$$

where $\omega_0 = 2\pi/\lambda_0$ is the wavenumber corresponding to the profile length λ_0 .

The fractal approaches to description of the surface roughness are caused by experimental observations that graphs of the spectral density $G(\omega)$ of the surface topography have often the power-law character

$$G(\omega) \sim 1/\omega^\psi \quad (1)$$

with any ψ in the range $1 < \psi \leq 3$; and one can expect that the fractal dimension of the graph $z(x)$ is equal to $(5 - \psi)/2$.

For synthesis of surfaces, one has to choose several characteristics whose values are compared with parameters of real surfaces. Since the common topographical characteristics of real surfaces are scale dependent parameters, the values of the chosen characteristics for synthetic surfaces have to be taken depending on the governing scale of the tribological process under consideration.

Key Applications

Recent developments in the synthesis of tribological surfaces have concentrated mainly on fractal approaches and wavelets. The popularity of fractal approaches is due to a common belief that fractal dimension of a surface is a scale-independent characteristic of roughness, while the wavelet approach is used due to its ability for surface synthesis independent of each specific scale.

Early Models of Rough Surfaces

In 1940 Zhuravlev introduced one of the first statistical models of roughness (Zhuravlev 2007). He assumed that (i) protuberances of the rough surfaces are spherical, (ii) their radii are the same and equal to R , (iii) they have various heights, and (iv) the distribution function of the number of protuberances at a specific height increases as the level of consideration goes deeper into the rough surface. The further development of this popular model is due to Greenwood and Williamson (1966). To employ Zhuravlev-Greenwood-Williamson type models, one has to define the radius of protuberances. If one assumes that the roughness is isotropic then the surface topography $z(x, y)$ is characterized by just a profile $z(x)$. If the profile is sampled at a regular spacing δx , i.e., one has the heights z_k , then the curvature (κ) of a protuberance z_k can be defined as

$$\kappa = -(z_{k-1} + z_{k+1} - 2z_k)/\delta x^2$$

where $z_{k-1}, z_{k+1} < z_k$ (Greenwood 1992). Whitehouse and Archard (1970) found that different sampling intervals gave different mean curvature, i.e., the mean curvature is a scale dependent parameter.

Another model of rough surfaces with spherical protuberances is due to Archard (1957). He covered

a sphere of radius R_0 by spherical protuberances of radius of curvature R_1 that are evenly distributed over the surface of the sphere. In the Archard model these protuberances in turn could be covered by smaller protuberances of radius R_2 ($R_2 \ll R_1 \ll R_0$). This model reflected a very important observation of multilevel structure of the roughness. Thus, synthetic surfaces normally have to have a multilevel structure.

Fourier and Wavelets Syntheses of Surface Topography

Fourier Synthesis

For analysis of surface topography, it is common to initially perform Fourier analysis of the surface shape, i.e., the surface data is decomposed using a complete orthogonal set of sine and cosine functions. The corresponding operation of rebuilding the topography up to the prescribed maximum number n_{max} of harmonics is known as synthesis. The case when the wavelength is much larger than the sample size is called long wavelength. Surface form is described by the longest spatial wavelength and defines the interval on which the surface is constructed. This is normally considered as the nominal shape of the surface. The first order (or harmonics) of the Fourier expansion is used to reconstruct this feature. Surface waviness represents the set of features from which the surface, at certain scale, does not behave in accordance with the nominal characteristic or shape. Roughness is sometimes considered as the noise of the surface. Thus, the long wavelength shapes are called "waviness," while the short wavelength features represent "roughness." In Fourier synthesis, the synthetic surface can be represented as

$$z(x) = \sum_{n=1}^{n_{max}} z_n \sin\left(\frac{2\pi nx}{2L} + \psi_n\right)$$

where n_{max} is the maximum number of waves, ψ_n are random phases, and for engineering surfaces typically $z_n = z_1/n$ (Morales-Espejel et al. 2000).

The main drawback of the Fourier synthesis is the following. Real profiles (surfaces) have a bounded domain where they are defined, while the Fourier transformation assumes that a surface can be covered by a set of trigonometric functions, which are defined on an infinite domain. This means that the analyzed synthetic surface has to be considered periodic. For this reason, it is possible to generate artifacts in the analysis of surfaces with non-periodic features, like spikes or cracks generated during a tribological experiment.

Wavelet Approach

The above described Fourier analysis is just one special case of spectral decomposition such that a surface profile is decomposed using trigonometric functions as a basis. Although this set of functions is very convenient from a mathematical point of view, other sets of waves are possible for decomposition of topography. One can take a function (prototype) such that its amplitude starts out at zero, changes its value, and then goes back to zero, to obtain a basis using dilations and translations of this function. The initial wavelet is called the mother (prototype) wavelet, while its scaled and translated copies are called daughter wavelets. An example of a wavelet is the Haar wavelet (Daubechies 1992),

$$\psi(x) = \begin{cases} 1 & 0 \leq x < \frac{1}{2} \\ -1 & \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

A decomposition of a one-dimensional profile using such non-trigonometric functions is called wavelet transform. One can assign a frequency range to each scaled copy (to each daughter wavelet). For a roughness profile, the wavelet transform gives a two-dimensional expansion of the profile with the frequency (scale) and the coordinate (location) treated as independent variables. Each scale component can be studied with a resolution that matches its scale.

Although the main idea of the wavelet transform (WT) is similar to the Fourier analysis, it has several distinctions. The main one is that the WT is constructed using bases (wavelets) having compact supports, i.e., each wavelet has a compact domain where it has non-zero values. The formula for a discrete wavelet transform is the following one (Daubechies 1992)

$$f = \sum_{m,n=-\infty}^{+\infty} c_{m,n} \psi_{m,n}$$

where

$$\psi_{m,n} = a_0^{-\frac{m}{2}} \psi(a_0^{-m}x - nb_0),$$

$\psi(x)$ is the mother wavelet, $c_{m,n}$ are the wavelet coefficients, a_0 is the coefficient that describes scale transformation (dilation), i.e., it defines the width of the wavelet, and b_0 is the coefficient that describes translations, i.e., it defines the wavelet position. Hence, the terms a_0 and b_0 represent, respectively, the expansion/shrinking and the translation of the mother wavelet $\psi(x)$. The wavelets must satisfy constraints that follow from the orthonormality of the basis.

Many researchers refer to wavelet analysis as a “mathematical microscope” (Astaf’eva 1996).

The principle is to consider each basis function as a mathematical probe on the given profile (surface). The coefficients a_0 and b_0 grant the possibility to move our mathematical probe along the surface (b_0) and to test the surface at different resolutions (a_0). For both computational reasons and easier comprehension, the dyadic wavelet form ($a_0 = 2$, $b_0 = 1$) has become the most common (Daubechies 1992). The choice of the mother wavelet depends on the type of feature under consideration.

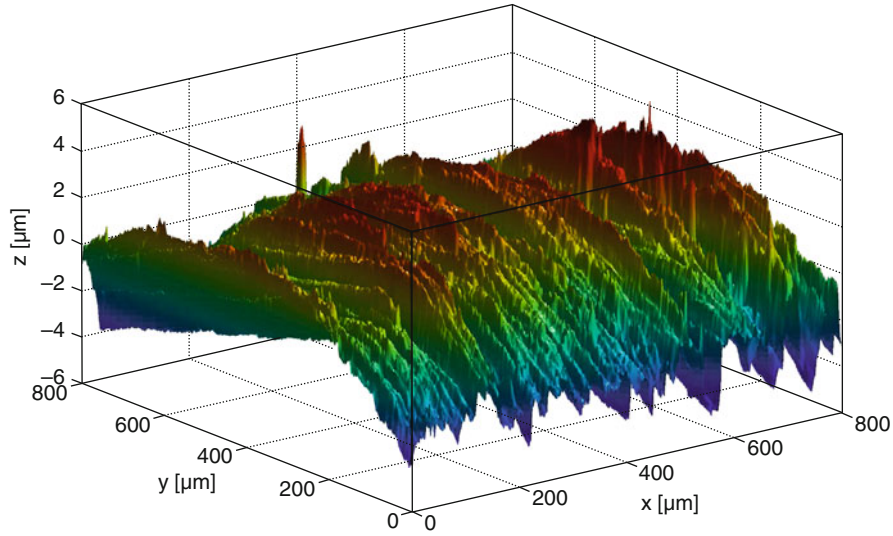
The advantage of the wavelet transform is caused mainly by the above mentioned factors: (i) the expansion basis functions (wavelets) have compact support; and (ii) the expansion coefficients are related simultaneously to the scale of consideration and to the location of the studying region of the surface. If one synthesizes a surface within a bounded domain, it is assumed that no information is given about the surface outside of the domain borders. Since wavelets have compact supports, the effect of the border will be reflected in the synthetic surface only within the narrow areas near the surface boundaries.

Example. A WT is able to identify local roughness on a surface. A sample of a measured or simulated surface of a grinded cylinder is shown in Fig. 1. Grinding causes non-periodic sharp discontinuities directed along y direction of the surface.

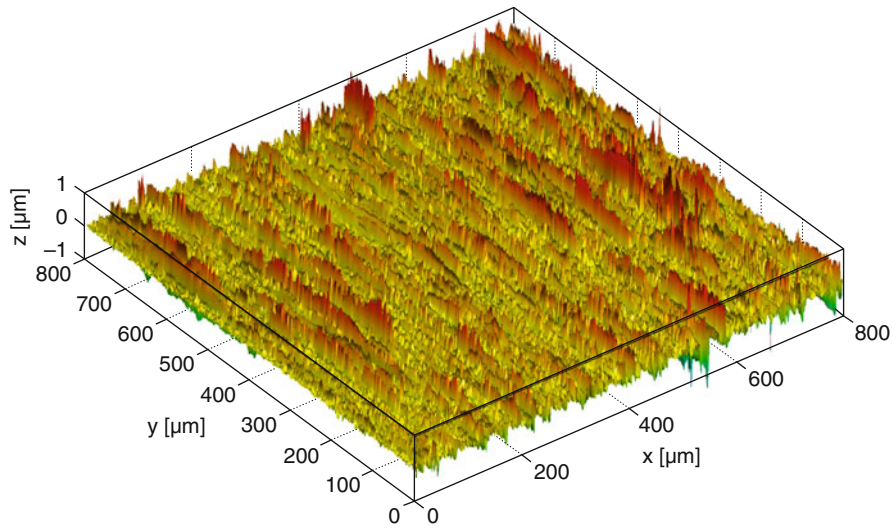
WT allows the researchers to separate the roughness texture from the surface and to study how the roughness changes locally. For this purpose, one can use a 2D wavelet transform whose coefficients may be obtained by application of the 1D WT in two directions (x and y). Figure 2 shows the texture of roughness that has been calculated by the proper selection (Stout and Blunt 2000) of the wavelet coefficients. One can see that WT has advantages over traditional Fourier analysis, because the step structure of the Haar wavelet allows the researchers to synthesize accurately the discontinuities and sharp peaks that are non-periodic along the x direction. For synthesis of local features of the surface of interest, one can generate easily a set of scale coefficients that define the roughness feature at different scale and invert the process in order to define a rough surface.

Weierstrass-Mandelbrot Profile

In 1986 Roques-Carmes et al. (1988) started to use Weierstrass type functions for fractal modeling of surface roughness. Then a number of other researchers modeled roughness using these functions. The Weierstrass-Mandelbrot (WM) function $W(x;p)$ depends on the scaling parameter p and fractal dimension D



Surface Synthesis Based on Surface Statistics, Fig. 1 Grinded cylinder with sharp discontinuities directed along y direction



Surface Synthesis Based on Surface Statistics, Fig. 2 Synthesized roughness matching the discontinuities of the grinded surface

$$W(x;p) = \sum_{n=-\infty}^{\infty} p^{(D-2)n} (1 - \cos p^n x).$$

The following truncated WM (\tilde{W}) function

$$\tilde{W}(x;p) = \Lambda a^{(D-1)} \sum_{n=n_1}^{\infty} p^{(D-2)n} \cos 2\pi p^n x,$$

$$1 < D < 2, \quad p > 1$$

was suggested for modeling any surface topography (Majumdar and Bhushan 1990). Here n_1 is an integer number, which corresponds to the low cut-off frequency of the profile, and Λ is the so-called characteristic length scale of the profile. The number n_1 depends on the length L of the sample and is given by $p^{n_1} = 1/L$ and the parameter Λ determines the position of the spectral density along the power axis. It was suggested to use the graph of \tilde{W} as synthetic profile whose fractal dimension

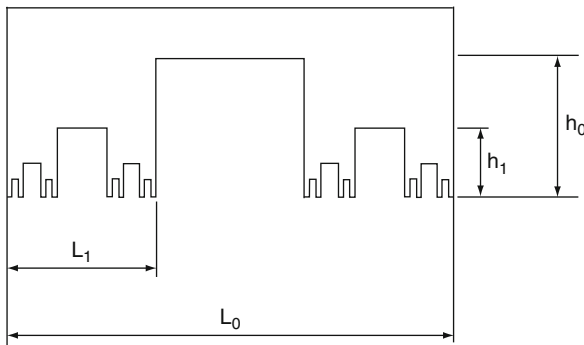
(the power spectra having a power-law fractal behavior) is the same as a real surface has. It was often stated that both parameters Λ and D of the function $\tilde{W}(x;p)$ are scale-invariant characteristics of the roughness. However, later the extensive experimental studies of this fractal characterization model showed that the values of parameters λ and D are not unique and depend on instruments or resolution of a given instrument (Bhushan 1995).

Further, the graph of $W(x;p)$ has a specific trend x^{2-D} . Hence, it can be used only as a particular example of a fractal profile and it cannot be considered as the general fractal functional model. The assumption that the WM function gives a universal fractal representation to all rough profiles can lead to wrong conclusions concerning surface roughness parameters and their distributions.

The C_B Profile and its Modifications

Another example of synthetic fractal surface is the Cantor-Borodich (C_B) profile. The procedure of construction of the C_B profile is quite similar to the construction of the Cantor-Liu bar model, however, the height of the C_B profile is always bounded. The original C_B profile has two scaling parameters f_x and f_z such that the width L_i and the height h_i of each generation of asperities are given by $L_i = L_0 f_x^i$ and $h_i = h_0 f_z^i$, respectively (Fig. 3). Although the C_B profile may be fractal for some values of the parameters $0 < f_x < 0.5$ and $0.5 < f_z < 1$, it is quite clear that it does not describe the roughness of a real solid. This model was introduced because it is possible to find analytical solutions to the contact problems for the C_B profile (Borodich and Onishchenko 1993).

Since the C_B profile does not have enough parameters to describe real rough surfaces, Warren and Krajcinovic (1996) suggested modifying the C_B profile and using



Surface Synthesis Based on Surface Statistics, Fig. 3 The original C_B profile with $s = 2$

a parameter s additionally to the scaling parameters f_x and f_z , where $s, s \geq 2$ is an integer that denotes the number of segments on a repeating surface. Hence, the width L_i and the height h_i of each generation of asperities are given by $L_i = L_0 (f_x/s)^i$ and $h_i = h_0 f_z^i$, respectively (Fig. 4). The original C_B profile corresponds to $s = 2$. The Minkowski (box) fractal dimension (D , $1 < D < 2$) of the modified C_B profile is

$$D = 1 - \frac{\ln f_z}{\ln s f_x} + \frac{\ln s}{\ln s f_x}.$$

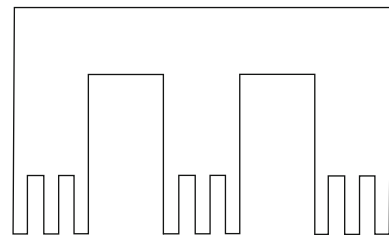
Later various other modifications of the C_B profile were introduced. However, all of these models have the same minor drawback: all asperities of the profiles have the one-level character, while as Archard showed that the structure of the real roughness is of hierarchical nature.

Some Other Synthetic Fractal Models

There are various other models of fractal surfaces, e.g., profiles based on Borodich's parametric-homogeneous functions (Borodich 1998), the fractional Brownian motion (fBm) models, and the spectral synthesis model (Glover et al. 1998). However, one has to realize that fractal dimension alone cannot give a full description of surface roughness. The above mentioned fractal models of surfaces can represent surfaces with a prescribed value of the fractal dimension, but they do not have a sufficient number of parameters to produce a surface with both prescribed engineering characteristics, and the fractal dimension. In addition, it is difficult to handle surface models based on the fBm profiles analytically, which presents a further obstacle in their study.

Multilevel Prefractal Model (the B-O Model)

Borodich and Onishchenko (1993) introduced an alternative to the C_B profile model, namely a multilevel model (the B-O model) of prefractal surfaces based on iterated function algorithms. It was noted that models



Surface Synthesis Based on Surface Statistics, Fig. 4 The modified C_B profile with $s = 3$

incorporating multilevel structures, when the asperities of the next generation lie on the tops of those of the previous level or when the even and odd generations are directed upwards and downwards correspondingly, could give more natural rough profiles than the C_B profile.

The B-O multilevel prefractional model was introduced as a fractal realization of Archard's idea of hierarchical structure of the roughness. The B-O model generates a hierarchical, prefractional curve profile when the asperities of the next generation lie on the tops of those of the previous level. The construction of iterated function models involves a sequence of approximations (prefractals), each obtained from its predecessor by modification in increasingly fine detail. The B-O model is characterized by two scaling structural parameters: horizontal factor α , and vertical factor β . Their meanings are clear from Fig. 5 showing a sketch of the MLP after the first three construction steps.

Consider a segment of length l_0 located along an axis Ox with the point O in the middle (see Fig. 5). Take l_0 as the base length of the B-O profile, and some value h_0 that can be easily connected with the total height H^* of the profile, and consider a block B_0 with length l_0 and height h_0 . This is an initial (zero) element for constructing of the B-O structure, which is not itself included in the structure. This object is called an *initiator*.

The B-O model can be constructed in a step-by-step manner. On the first step, two positive structural parameters of the model α and β are fixed. The proposed model can be both fractal and non-fractal depending on values of the structural parameters. Regardless of this, the model profile remains rough and possesses certain self-affine (two-dimensional scaling) properties. The iterative regular construction of the profile allows the researchers to analyze its prestructures (prefractals) of arbitrary generation. The parameters should satisfy the following conditions

$$0 < \alpha < 1/2, \quad 0 < \beta < 1 \quad (3)$$

in order to produce a fractal profile. The horizontal interval l_0 is split in three subintervals whose lengths are

equal to αl_0 , $(1 - 2\alpha)l_0$ and αl_0 , respectively. Then a block of the first generation having height βh_0 is built above the middle subinterval. This is the approximation of the B-O model of the first generation.

On the second step, the above described procedure is repeated for each of the three horizontal subintervals. Note the middle subinterval is located on the height βh_0 . After this step, there are nine horizontal subintervals located at different heights.

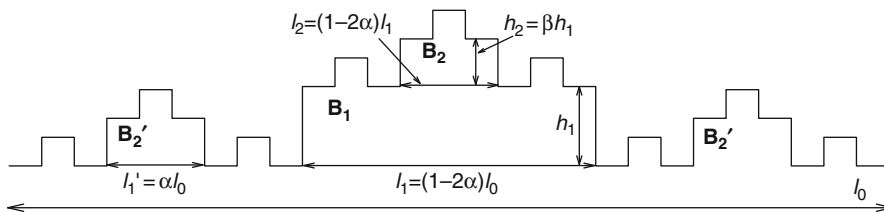
The described procedure is repeated up to a given number of generations.

It can be shown (Borodich and Onishchenko 1999) that fractal (box) dimension of the B-O profile in the particular case $\alpha = 1/3$ is given by

$$d_B = 2 - \frac{\ln\left(\frac{1}{\beta}\right)}{\ln 3}.$$

The multilevel models have several important advantages in comparison with C_B and other synthesized fractal curves that are currently used for modeling of rough surfaces. In particular, the inherent multilevel structure of the models allows the researchers the flexible separation of macro-, micro-, and nano- (molecular) roughness as well as to reflect other scale-oriented types of roughness. In addition, although the shape of the modeled surfaces appears jagged and irregular, it is completely regular from the viewpoint of the construction algorithm.

The fractal dimension does not reflect details of surface manufacturing and it does not characterize the contact properties of the surfaces. However, if the considered tribological process depends on the roughness having the power spectral density of the form (1) within some bounded range, then the synthetic surface has to have the same prefractional features. One may expect that the B-O model can model surfaces with prescribed values of several statistical tribological characteristics of the real surfaces and additionally it may have fractal-like behavior in the bounded interval of scales.



Surface Synthesis Based on Surface Statistics, Fig. 5 The iterative construction algorithm for the B-O multilevel prefractional curve

Cross-References

- [Contact of rough surfaces: The Greenwood and Williamson/Tripp, Fuller and Tabor Theories](#)
- [Fractal Characterization of Surfaces](#)
- [Fractal Contact Mechanics](#)
- [Fractal Geometry](#)
- [Fractal Nature of Surfaces](#)
- [Surface Characterization and Description](#)

References

- J.F. Archard, Elastic deformation and the laws of friction. *Proc. R. Soc. Lond.* **A243**, 190–205 (1957)
- N.M. Astaf'eva, Wavelet analysis: basic theory and some applications. *Phys. Usp.* **39**, 1085–1108 (1996)
- B. Bhushan, A fractal theory of the temperature distribution at elastic contacts of fast sliding surfaces – discussion. *J. Tribol.* **117**, 214–215 (1995)
- F.M. Borodich, Parametric homogeneity and non-classical self-similarity. II. Some applications. *Acta Mech.* **131**, 47–67 (1998)
- F.M. Borodich, D.A. Onishchenko, Fractal roughness for problem of contact and friction (the simplest models). *J. Fricti. Wear* **14**, 452–459 (1993)
- F.M. Borodich, D.A. Onishchenko, Similarity and fractality in the modelling of roughness by a multilevel profile with hierarchical structure. *Int. J. Solids Struct.* **36**, 2585–2612 (1999)
- I. Daubechies, *Ten Lectures on Wavelets*. CBMS-NSF Lecture Notes, vol. 61 (SIAM, Philadelphia, 1992)
- P.W.J. Glover, K. Matsuki, R. Hikima, K. Hayashi, Synthetic rough fractures in rocks. *J. Geophys. Res. Solid Earth* **103**(No.B5), 9609–9620 (1998)
- J.A. Greenwood, Problems with surface roughness, in *Fundamentals of Friction: Macroscopic and Microscopic Processes*, ed. by I.L. Singer, H.M. Pollock (Kluwer, Boston, 1992), pp. 57–73
- J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces. *Proc. R. Soc. Lond.* **A370**, 300–319 (1966)
- A.A. Lubrecht, C.H. Venner, Elastohydrodynamic lubrication of rough surfaces. *Proc. Inst. Mech. Eng. Pt. J J. Eng. Tribol.* **213**, 397–404 (1999)
- A. Majumdar, B. Bhushan, Role of fractal geometry in roughness characterization and contact mechanics of surfaces. *J. Tribol.* **112**, 205–216 (1990)
- G.E. Morales-Espejel, C.H. Venner, J.A. Greenwood, Kinematics of transverse real roughness in elastohydrodynamically lubricated line contacts using Fourier analysis. *Proc. Inst. Mech. Eng. Pt. J J. Eng. Tribol.* **214**, 523–534 (2000)
- B. Nowicki, Multiparameter representation of surface roughness. *Wear* **102**, 161–176 (1985)
- C. Roques-Carmes, D. Wehbi, J.F. Quiniou, C. Tricot, Modelling engineering surfaces and evaluating their non-integer dimension for application in material science. *Surf. Topogr.* **1**, 435–443 (1988)
- K.J. Stout, L. Blunt, Filtering technology for three-dimensional surface topography analysis, in *Three Dimensional Surface Topography*, ed. by K.J. Stout, L. Blunt (Penton Press, London, 2000), pp. 95–142
- T.L. Warren, D. Krajcinovic, Random Cantor set models for the elastic perfectly plastic contact of rough surfaces. *Wear* **196**, 1–15 (1996)
- D.J. Whitehouse, J.F. Archard, The properties of random surfaces of significance in their contact. *Proc. R. Soc. Lond.* **A 316**, 97–121 (1970)
- V.A. Zhuravlev, On question of theoretical justification of the Amontons-Coulomb law for friction of unlubricated surfaces. *Proc. Inst. Mech. Eng. Pt. J J. Eng. Tribol.* **221**, 397–404 (2007)

Surface Temperature for Sliding Surfaces

- [Contact Temperature of a Moving Solid Surface](#)

Surface Temperature for Stationary Surfaces

- [Contact Temperature of a Stationary Solid Surface](#)

Surface Temperature Rise of EHL Contacts

- [Flash Temperature in EHL Contacts](#)

Surface Temperatures in Dry Sliding

- [Frictional Heat Generation, Partitioning, and Dissipation in Dry Tribological Contacts](#)

Surface Tension

- [Surface Forces, Surface Tension, and Adhesion](#)

Surface Tension and Adhesion

- [Basic Concepts in Adhesion Science](#)

Surface Tension and Surface Free Energy

- [Surface Free Energy](#)

Surface Tension in Biosystems

► Adhesion in the Animal World

Surface Texture for Bodies in Non-Conformal Contacts

IVAN KRUPKA

Brno University of Technology, Brno, Czech Republic

Synonyms

Surface texturing for non-conformal surfaces; Surface topography modification

Definition

Surface topography modification using surface features of appropriate size and shape to form micro-reservoirs emitting additional lubricant to the non-conformal contact between rubbing surfaces in relative motion

Scientific Fundamentals

Surface Texturing

The surface texturing approach is relatively straightforward. Surface features (micro-cavities) are introduced onto the surface as lubricant micro-reservoirs. The lubricant is emitted at their downstream and upstream sides under rolling/sliding conditions when the micro-reservoir is on the slower or faster moving surface, respectively (Fig. 1). It is caused by the lubricant traveling through the contact with the average speed of the surfaces so that lubricant entrapped within the micro-cavity goes ahead of it or lags behind it and elastically deforms the rubbing

surfaces (Ai and Cheng 1994). In such a way, by using an array of appropriate micro-cavities, the additional lubricant supplied to the non-conformal contact helps to separate rubbing surfaces.

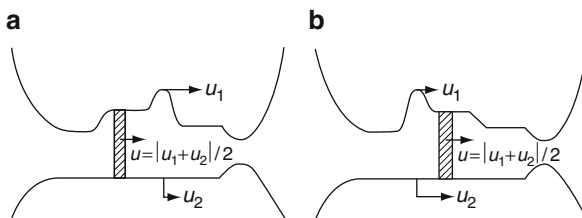
Effect of Micro-cavities on Film Thickness

However, the effect of such micro-cavities on non-conformal lubricated contact is much more complex and is strongly influenced by their size (especially depth) and shape. An increase in film thickness is accompanied by a reduction in film thickness that is caused by a notable side leakage of the entraining fluid as well as a fall of hydrodynamic pressure while the micro-cavity is entering the contact (Kaneta et al. 1997). The film thickness reduction locally extends upstream or downstream in the form of two thin strips along the thick film region under rolling/sliding conditions. Figure 2 depicts such an effect observed within lubricated contact formed between a glass disk and steel ball with the single dent (1 μm in depth) on its surface. Such a film thickness reduction increases with increasing size (depth) of micro-cavity and can even result in lubrication film breakdown.

Side leakage becomes a significant factor mainly for large micro-features. Micro-cavities located close to the contact border can cause significant film thickness reduction due to the side leakage of the lubricant. Therefore, transversally or longitudinally oriented grooves are not suitable for surface texturing applications (Krupka and Hartl 2009). From this point of view, surface texturing based on small micro-features of circular shape should be considered (Fig. 3). Moreover, the depth of micro-cavities should be selected carefully to ensure that the beneficial effect of enlarged film thickness is not outweighed by possible film thickness collapse.

Effect of Micro-cavities on Pressure

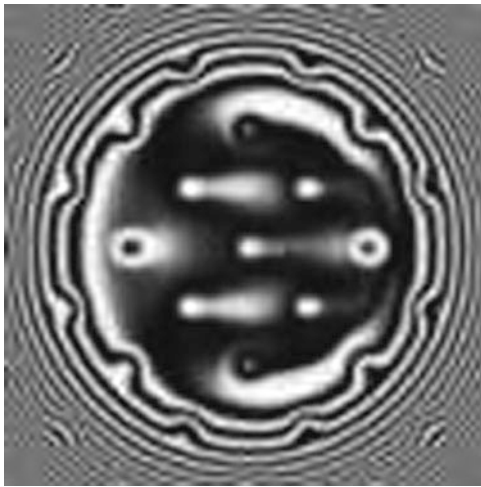
The presence of micro-cavities within highly loaded contacts results not only in significant changes in lubricant film thickness but also in pressure distribution. Highly localized pressure peaks in their vicinity increase subsurface stresses that may eventually initiate microcracks on the surface leading to fatigue failure. Figure 4 shows that, under elastohydrodynamic lubrication conditions, even shallow micro-cavities with a depth of 175 nm increase the contact pressure significantly. However, while micro-features introduced to the rubbing surfaces can decrease the life of machine elements under full film lubrication, quite a different effect can be considered with their use under mixed lubrication conditions. Asperity interactions result in significant variation of pressure field within the contact, and fatigue life is significantly reduced. In this



Surface Texture for Bodies in Non-Conformal Contacts,
Fig. 1 Lubricant emitted from a surface micro-cavity with the textured surface moving faster (a) or slower (b) than the smooth surface



Surface Texture for Bodies in Non-Conformal Contacts, Fig. 2 Effect of a surface dent on film thickness (in μm) for $u_{\text{ball}} = 3u_{\text{disk}}$ (a), $u_{\text{disk}} = u_{\text{ball}}$ (b) and $u_{\text{disk}} = 3u_{\text{ball}}$ (c) (Kaneta et al. 1997)



Surface Texture for Bodies in Non-Conformal Contacts, Fig. 3 Transient increase of film thickness in a micro-textured contact (Mourier et al. 2006)

case the application of surface texturing can be considered a way to decrease the number of asperity interactions and thereby reduce the risk of rolling contact fatigue (Akamatsu et al. 1992). The micro-cavities have to be much deeper (at least several micrometers) compared with those suggested for surface texturing applications to cause a noticeable reduction in contact fatigue life under mixed lubrication conditions (Krupka et al. 2008). Rolling contact fatigue tests showed that fatigue life was significantly reduced when a dent with a depth about $20\text{ }\mu\text{m}$ was located within the track. Conversely, no influence on rolling contact fatigue due to the dent was observed when the depth of the dent was reduced to $4\text{ }\mu\text{m}$, which is well above the depth considered for surface texturing applications within non-conformal contacts. This is why,

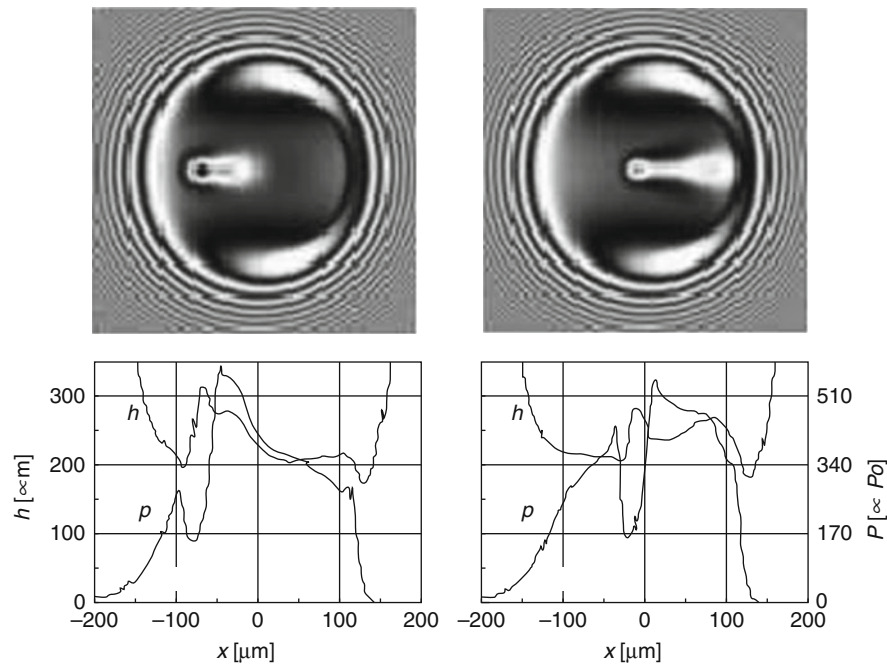
under mixed lubrication conditions, the effect of surface micro-cavities on film thickness represents more critical parameter and limits the use of surface texturing to shallow micro-cavities having a depth of several hundred nanometers.

Surface Texture Creation

The effect of micro-cavities' shape and size on film thickness and pressure distribution within non-conformal contacts represents the limits for their practical use in surface topography modification. Micro-cavities should be small (comparing the contact dimensions) and shallow and without the lip of raised material localized around the micro-cavity perimeter. These requirements are rather restrictive, especially for industrial applications. In laboratory conditions such micro-cavities can be prepared by femtosecond laser (Fig. 5) or ion beam ablation or by a diamond indenter that is hardly applicable in a mass production. The application of surface texturing in highly loaded machine elements would require the development of effective technique for micro-cavities production. Nevertheless, the selection of appropriate surface finishing techniques producing surface topography containing roughness features working as surface micro-cavities can represent perspective approach. Numerical simulations can be involved into design engineering surfaces through surface optimization, taking into account both the surface topography and lubrication performance (Wang and Zhu 2005). The development of adequate three-dimensional characterization and description of rubbing surfaces is required (Stachowiak and Podsiadlo 2008).

Key Applications

The complex effect of surface features on film thickness and pressure distribution within non-conformal contacts is principal for possible surface texturing applications.



Surface Texture for Bodies in Non-Conformal Contacts, Fig. 4 Effect of micro-cavity passing through a circular EHD contact on film thickness and pressure distribution (Mourier et al. 2006)

It requires well-balanced design taking into account many factors, including lubrication regime, slide-to-roll ratio conditions, rolling contact fatigue, and technology of surface texturing patterns creation. The application of surface texturing to the highly loaded non-conformal contacts operated under full film or boundary lubrication conditions appears to be questionable because of the increasing risk of lubrication breakdown or fatigue failure. In these cases lubricant chemistry represents the optimal alternative. Conversely, many machine components are operated under the conditions when the rubbing surfaces are not completely separated by the lubrication film. In such a case surface micro-cavities can help reduce friction and wear of rubbing surfaces.

Mixed Lubrication

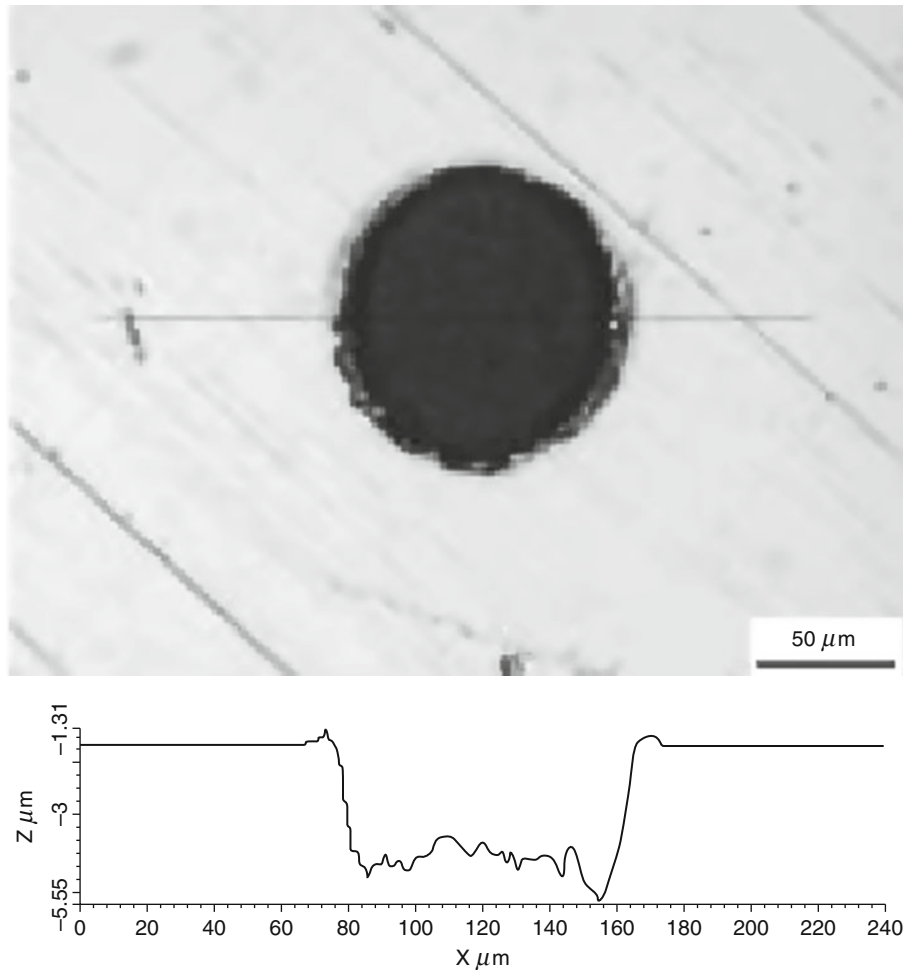
Surface topography significantly influences the behavior of mixed lubricated contacts where rubbing surfaces are only partially separated by lubrication film. Introduction of surface micro-cavities to the surfaces should help to reduce asperity interactions and thereby reduce friction and wear of rubbing surfaces. Figure 6 shows the effect of

surface micro-cavities on a mixed lubricated contact formed between a textured steel ball and a smooth glass disk. The ball surface is indented using a Rockwell indenter. The distance between adjoining dents is 50 μm and their depth is between 200 and 300 nm. It can be seen from chromatic interferograms and film thickness profiles that lubricant emitted by micro-dents can effectively lift off the real roughness features and thereby reduce asperity interactions. Lubricant emitted from surface features appears to be spread more easily within the mixed lubricated contact because of local decreases of pressure in the vicinity of real surface features. The presence of shallow surface features can help to separate mixed lubricated rubbing surfaces more efficiently than can be suggested from the results obtained with smooth surfaces.

Start-up Conditions

The start-up operation of non-conformal contacts represents another case that risks surface damage because of asperity interactions.

Numerical simulation of a non-conformal line contact with a stationary central surface pocket showed that the

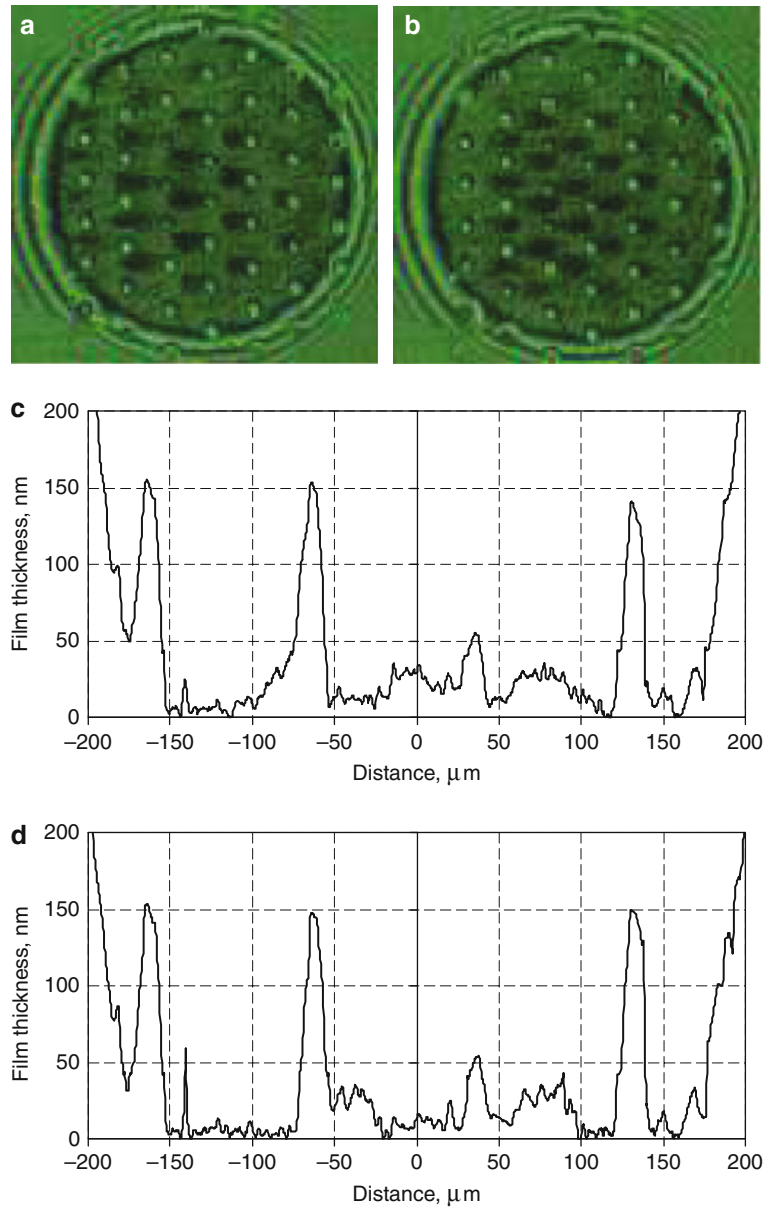


Surface Texture for Bodies in Non-Conformal Contacts, Fig. 5 Micro-cavity produced by femtosecond laser (200 pulses -0.71 J/cm^2) (Mourier et al. 2008)

lubricant trapped inside a pocket can positively influence start-up behavior (Zhao and Sadeghi 2004). Compared with a smooth surface start-up condition, there is less frictional heat generated in the contact area and thus the surface temperature rise during the start-up process is much smaller. These effects can be helpful in applications with frequent start and stop or oscillating running conditions, in which direct solid to solid contact occurs at the start (or restart) of motion.

The beneficial effect of surface texturing on non-conformal contact during start-up is shown in Fig. 7, which compares lubrication film formation with textured and non-textured contact. The distance between adjoining micro-cavities is $50 \mu\text{m}$ and their depth is about 200 nm .

It can be seen that during start-up motion the lubricant squeezed within surface micro-cavities is emitted downstream of the surface features. It enlarges film thickness and reduces surface interactions as can be seen qualitatively on chromatic interferograms as darker spots downstream of the micro-cavities. In Fig. 7b, however, there are also such dark spots moving through the contact area where the rubbing surfaces are still in solid-to-solid contact. These darker spots again correspond to the enlarged film thickness regions as a result of the lubricant emitted from the shallow pits produced on the rubbing surface during the surface finishing process. This demonstrates that proper selection of a surface finishing procedure can produce an alternative to surface texturing.



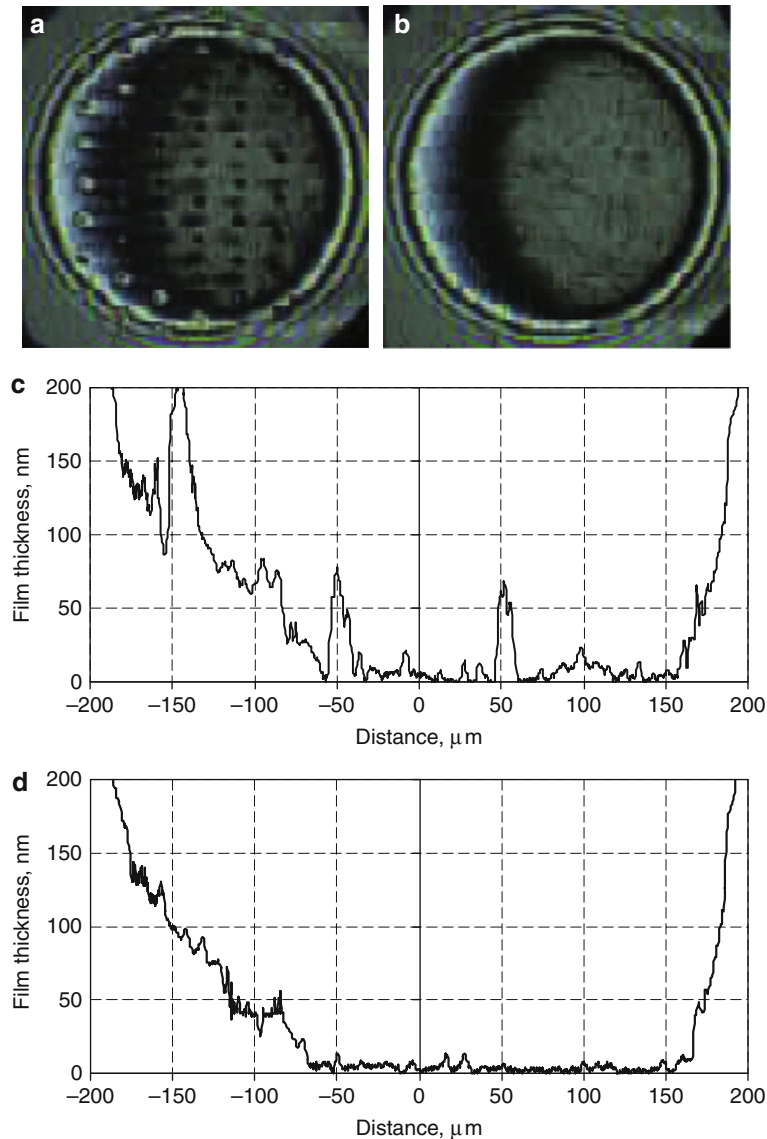
Surface Texture for Bodies in Non-Conformal Contacts, Fig. 6 Effect of surface micro-cavities on mixed lubricated contact with the textured surface moving faster (a, c) or slower (b, d) than the smooth surface

Starvation

Poor lubrication conditions represent another possible application for surface texturing. An insufficient supply of lubricant may result in starvation and consequently in

lubrication film collapse. Surface micro-cavities can be used to increase the supply of lubricant to the contact.

While the micro-cavity filled with lubricant is entering the contact, part of its content is released in the inlet zone.

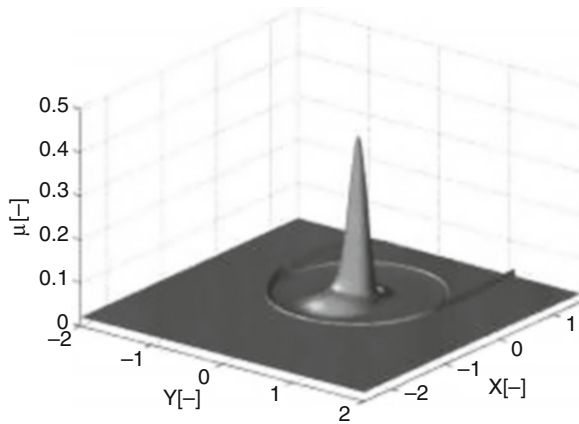


Surface Texture for Bodies in Non-Conformal Contacts, Fig. 7 Lubrication film formation during start-up with textured (**a, c**) and non-textured surfaces (**b, d**)

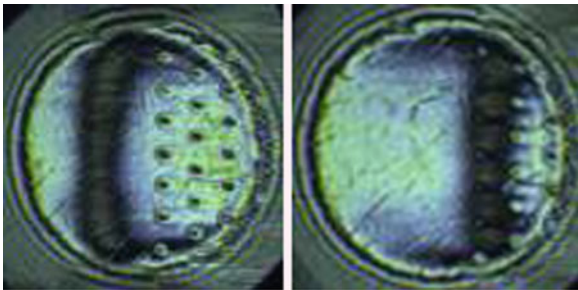
A fraction of this released oil remains in the inlet region and produces an upstream shift of the meniscus (Dumont et al. 2002). The remainder of the oil is entrained and builds up the film thickness at the trailing edge of the micro-cavity. Figure 8 shows such an effect on lubrication film within non-conformal contact using a micro-cavity with the depth of 1 μm. In this case the beneficial effect is obtained under pure rolling conditions, unlike the other

surface texturing applications where some sliding is required.

Surface micro-cavities can help to separate rubbing surfaces within starved non-conformal contact not only by supplying more lubricant to the contact entrance. Additional lubricant can be emitted from micro-cavities to the contact (Fig. 9) that locally enlarges the separation of rubbing surfaces under rolling/sliding conditions.



Surface Texture for Bodies in Non-Conformal Contacts,
Fig. 8 Effect of surface micro-cavity on lubrication film within
 a severely starved contact (Dumont et al. 2002)



Surface Texture for Bodies in Non-Conformal Contacts,
Fig. 9 Effect of surface micro-cavities on lubrication film
 within a starved contact

Cross-References

- [EHL Film Thickness Behavior](#)
- [Elastohydrodynamic Lubrication](#)
- [Friction/Traction Behavior of EHL](#)

References

- X. Ai, H.S. Cheng, The influence of moving dent on point EHL contacts. *Tribol. Trans.* **37**(2), 323–335 (1994)
- Y. Akamatsu et al., Influence of surface roughness skewness on rolling contact fatigue life. *Tribol. Trans.* **35**, 745–750 (1992)
- M. Dumont et al., Surface feature effects in starved circular EHL contacts. *Trans. ASME, J. Tribol.* **124**, 358–366 (2002)
- M. Kaneta et al., *Optical Interferometric Observations of the Effects of a Moving Dent on Point Contact EHL*. In: *Elastohydrodynamics '96 Fundamentals and Applications in Lubrication and Traction*. Tribology Series 32, Elsevier, Amsterdam, 1997, pp. 69–79
- I. Krupka et al., Effect of surface texturing on mixed lubricated non-conformal contacts. *Tribol. Int.* **41**(11), 1063–1073 (2008)
- I. Krupka, M. Hartl, Effect of surface texturing on very thin film EHD lubricated contacts. *Tribol. Trans.* **52**(1), 21–28 (2009)
- L. Mourier et al., Transient increase of film thickness in micro-textured EHL contacts. *Tribol. Int.* **39**, 1745–1756 (2006)
- L. Mourier et al., Action of a femtosecond laser generated micro-cavity passing through a circular EHL contact. *Wear* **264**(5–6), 450–456 (2008)
- G. Stachowiak, P. Podsiadlo, 3-D characterization, optimization, and classification of textured surfaces. *Tribol. Lett.* **1**, 13–21 (2008)
- Q.J. Wang, D. Zhu, Virtual texturing: modeling the performance of lubricated contacts of engineered surfaces. *Trans. ASME, J. Tribol.* **127**(4), 722–728 (2005)
- J.X. Zhao, F. Sadeghi, The effects of a stationary surface pocket on EHL line contact start-up. *Trans. ASME, J. Tribol.* **126**, 672–680 (2004)

Surface Texture for Magnetic Recording

FRANK E. TALKE

Mechanical and Aerospace Engineering, University of
 California, San Diego, La Jolla, CA, USA

Synonyms

[Laser texturing](#); [Stiction in magnetic recording](#); [Tape rollers](#); [Textured slider](#); [Tribology of the head/disk interface](#)

Definition

Magnetic recording is the predominant technology for storing digital and analogue information on disks and tapes. It is accomplished by the relative motion between a read/write head and the magnetic medium. Surface texturing of hard disks can improve the tribology of the head/medium interface.

Scientific Fundamentals

The tribological performance of magnetic recording disk and tape drives can be enhanced by using mechanical or laser texturing of either the magnetic media or the magnetic read/write head. In disk drives, texturing has been used to improve the orientation of the grains in the magnetic media and also to provide a surface with desirable roughness properties so that the magnetic recording slider (head) can fly at a minimum flying height over the disk surface.

Key Applications

Texturing has been used to reduce stiction and friction in the landing zone of hard disks and to enhance the wear resistance of the slider/disk interface. In tape drives, texturing of guide rollers has been explored as a means for reducing the friction between magnetic tape and mechanical guide elements.

Principle of Magnetic Recording in Disk and Tape Drives

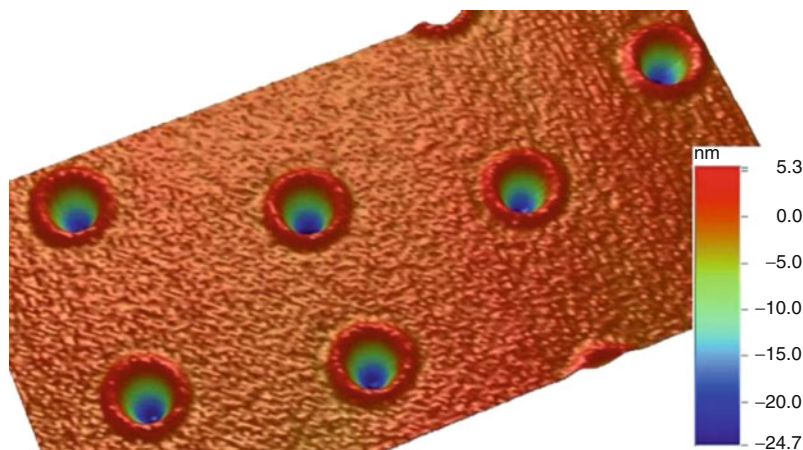
“Conventional” magnetic recording is accomplished by the relative motion between a magnetic medium (disk or tape) and a read/write transducer. In order to achieve maximum recording density, the spacing between the magnetic medium and the read/write transducer must be as small as possible without causing wear at the interface. In disk drives, the read/write transducer is incorporated in a so-called slider, which flies at close spacing over the disk surface. The spacing between the read/write transducer and the magnetic disk is currently less than 10 nm, requiring very “smooth” surfaces on the disk and the slider. Typical values of root mean square roughness of present-day disks are on the order of 0.25 nm, while the peak-to-valley spacing is less than 2 nm. These low values of surface roughness are likely to cause stiction between the slider and the disk whenever the slider rests on the disk or whenever contacts occur between slider and disk. Stiction is undesirable as it causes material interactions and wear between the transducer and the disk.

Texturing of the Landing Zone of Thin-Film Disks

In order to reduce stiction during contact start stop, texturing of the landing zone of a hard disk drive has been implemented in contact start stop drives. Texturing is performed by “creating” evenly spaced “bumps” on the disk surface using a high-power pulsed laser. This so-called laser texturing is applied on the inner diameter

of the magnetic hard disk over a narrow band in the radial direction, the so-called landing zone, prior to the deposition of the magnetic thin film. Crater-like depressions with a surrounding circumferential ridge can be created by this process and have been shown to be highly effective in reducing stiction between the slider and the disk. The height of the ridges of typical laser-textured surfaces can be adjusted to be in the 2–20 nm range. During start/stop of the hard disk drive, the slider starts and lands on the laser bumps. This reduces stiction and wear at the contact area between slider and disk. The tribological performance of laser-textured disks (take-off distance, stiction, and wear) depends strongly on the shape and size of the laser bumps as well as the flying height and contact area between the slider and the disk (Knigge et al. 1998; Baumgart et al. 1995). Figure 1 shows an image of a typical laser-textured disk area measured with an optical profiler. The bump height is on the order of 5.3 nm.

The radius of curvature of the laser bumps has a large impact on the stiction and wear characteristics of the head/disk interface. If the height of the laser bumps is larger than the flying height of the slider, sliding contact and wear can occur leading to an early failure of the head/disk interface. On the other hand, if the height of the bumps is very small compared with the flying height of the slider, stiction during contact between slider and disk can increase. Thus, the flying height, bump height, and contact area between slider and disk must be optimized for each particular slider/disk combination. Typical laser bump cross sections for the sombrero-, volcano-, and w-type are shown in Fig. 2.



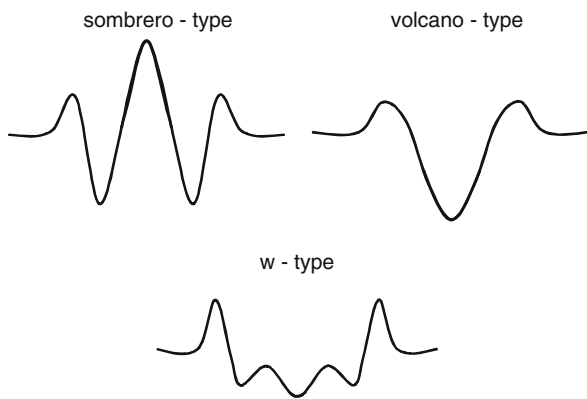
Surface Texture for Magnetic Recording, Fig. 1 Typical laser-textured hard disk area

Mechanical texturing of the landing zone of magnetic hard disks rather than laser texturing has also been attempted in the past. Mechanical texturing is more difficult to implement than laser texturing and has shown that stiction increases over time caused by an increase in the contact area due to smoothening of the asperities (Khurshudov et al. 1997). Figure 3 shows the stiction force as a function of contact-start-stop cycles for a slider moving on a mechanically and laser-textured surface, respectively. It can be seen that the stiction force

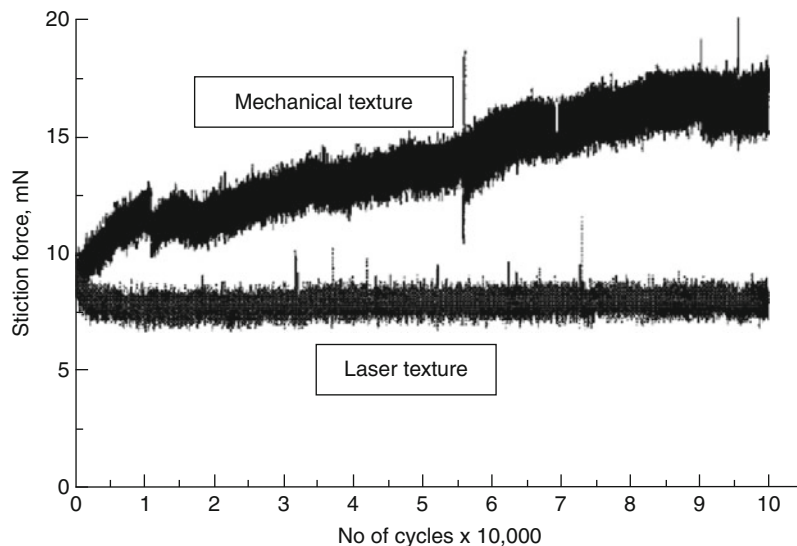
increases for mechanically textured surfaces, while it remains nearly constant for laser-textured media.

Texturing of Magnetic Recording Sliders

An alternate approach to reduce stiction between a slider and a disk is texturing of the slider instead of the disk. Texturing of sliders allows the use of very smooth disks, a prerequisite for achieving very small head/disk spacing. Magnetic recording sliders are made of alumina titanium carbide, $\text{Al}_2\text{O}_3\text{-TiC}$. If $\text{Al}_2\text{O}_3\text{-TiC}$ is exposed to reactive ion beam or oxygen plasma, different etching rates are observed for Al_2O_3 and TiC. This difference in the etching rate of Al_2O_3 and TiC can be used in the texturing of the air-bearing surface by creating a bimodal distribution of surface heights on the slider surface. Figure 4 shows an AFM image of a textured $\text{Al}_2\text{O}_3\text{-TiC}$ slider. Islands of TiC protruding above the recessed Al_2O_3 regions are clearly visible. Texturing of the slider reduces the friction due to a decrease in the area of contact between slider and disk. In addition, intermolecular forces, which are present at the low spacing between slider and disk, are decreased by this reduction of contact area as well. Texturing of sliders was also found to cause a reduction in the flying height as a function of the contour design of the slider air bearing (Zhang et al. 2005). Surface texture on magnetic recording sliders has also been found to reduce slider in-plane and out-of-plane vibration modes at ultralow-flying heights, i.e., flying heights below 10 nm.



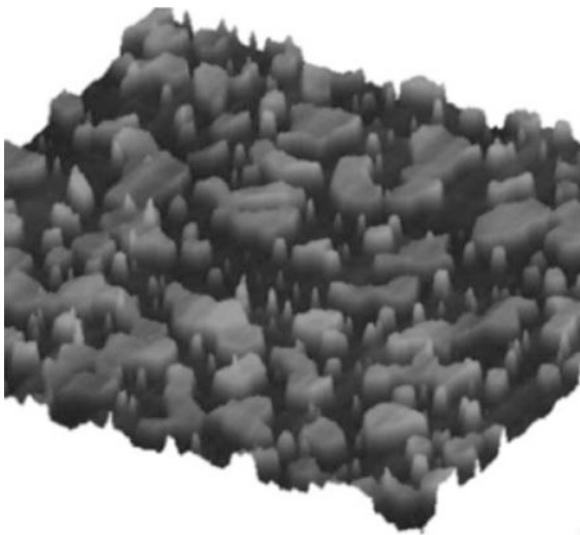
Surface Texture for Magnetic Recording, Fig. 2 Cross section of typical laser bump types



Surface Texture for Magnetic Recording, Fig. 3 Stiction force vs. number of CSS cycles for a slider on a mechanically and laser-textured surface (Khurshudov et al. 1997)

Mechanical Texturing of Disks to Improve Orientation of the Grains in the Magnetic Film

Mechanical texturing of magnetic disks can also be used to improve the magnetic orientation of the grains on the disk, thereby enhancing the magnetic performance of disk drives (Katayama et al. 1988). To align the orientation of the grains, a pattern of closely spaced circumferentially oriented “grooves” is made on the substrate prior to sputtering or chemical vapor deposition of the magnetic thin film. The grooves become sites for crystal growth of the magnetic material and influence the alignment of the magnetic grains.



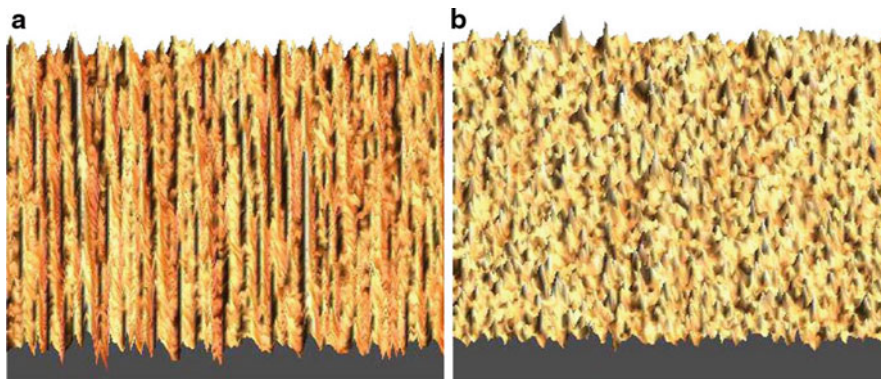
Surface Texture for Magnetic Recording, Fig. 4 AFM image of textured slider surface (Zhang et al. 2005)

Depending on the anticipated use of a disk drive, glass or aluminum substrates are used. In the case of aluminum disks, a nickel-phosphorous layer is applied to the bare disk to “hide” the alumina-magnesium substrate. Texturing of hard disks has been implemented in the form of circumferentially oriented texture or as isotropic texture. An image of a circumferentially textured disk and a randomly oriented textured disk is shown in Fig. 5 (Khurshudov and Raman 2005).

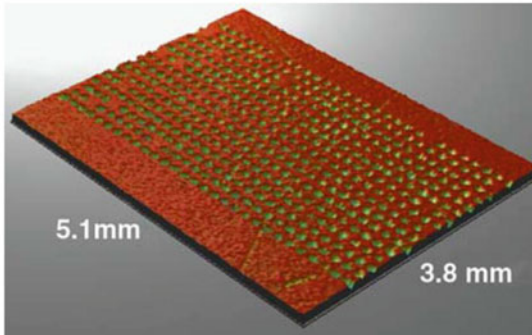
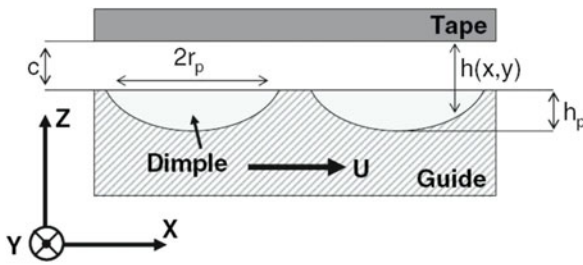
Texturing of the Tape/Guide Roller Interface in Magnetic Tape Drives

The friction between guide rollers and magnetic tape can be reduced by laser surface texturing of the cylindrical guides (Raeymaekers et al. 2007). Laser-textured surface features enhance the formation of an air bearing between the rollers and the tape and, therefore, reduce the coefficient of friction.

Laser surface texturing allows the formation of a hydrodynamic air bearing at low velocities and reduces the influence of speed on the friction coefficient. The critical speed where boundary lubrication changes into full fluid lubrication decreases significantly for laser-textured guides. The use of laser-textured guides has been suggested in tape drives using metal-evaporated (ME) tapes, since the latter tapes are magnetically superior to other types of tape but inferior in their tribological performance compared with metal particulate (MP) tape. At low speeds, the ME tape tends to “stick” to tape guides. However, the use of laser surface textured guides would improve the tribological performance of ME tapes in high-performance commercial drives by reducing stiction and wear (Fig. 6).



Surface Texture for Magnetic Recording, Fig. 5 Circumferential (a) and random (b) texture in hard disks (Khurshudov and Raman 2005)



Surface Texture for Magnetic Recording, Fig. 6 Geometry of dimple pattern created by laser surface texturing (Raeymaekers et al. 2007)

Cross-References

- [Adhesive Contact of Elastic Bodies](#)
- [Basic Concepts in Adhesion Science](#)
- [Bonded Solid Lubrication Coatings, Process and Applications](#)
- [Diamond-Like Carbon Coatings](#)
- [Fractal Characterization of Surfaces](#)
- [Homogenization of the Reynolds Equation](#)
- [Laser Texturing, Characterization and Related Effects](#)
- [Reynolds Equation](#)
- [Reynolds Equation for Compressible Fluid or Gas Film](#)
- [Surface Analysis Using Contact Mode AFM](#)
- [Surface Texture Generation with a Numerical Process](#)
- [Tribochemistry of Computer Disks](#)

References

- P. Baumgart, D. Krajnovich, T. Nguyen, A. Tam, A new laser texturing technique for high-performance magnetic disk drives. *IEEE Trans. Magn.* **31**(6), 2946–2951 (1995)
- S. Katayama, T. Tsuno, K. Enjoji, N. Ishii, K. Sono, Magnetic properties and read-write characteristics of multilayer films on a glass substrate. *IEEE Trans. Magn.* **24**(6), 2982–2984 (1988)
- A. Khurshudov, V. Raman, Roughness effects on head-disk interface durability and reliability. *Trib. Int.* **38**, 646–651 (2005)
- A. Khurshudov, B. Knigge, F. Talke, Tribology of laser-textured and mechanically-textured media. *IEEE Trans. Magn.* **33**(5), 3190–3192 (1997)

- B. Knigge, Q. Zhao, F. Talke, Tribological properties and environmental effects of nano and pico sliders on laser textured media. *IEEE Trans. Magn.* **34**(4), 1732–1734 (1998)
- B. Raeymaekers, I. Etsion, F. Talke, Enhancing tribological performance of the magnetic tape/guide interface by laser surface texturing. *Trib. Lett.* **27**(1), 89–95 (2007)
- J. Zhang, L. Su, F. Talke, Effect of surface texture on the flying characteristics of pico sliders. *IEEE Trans. Magn.* **41**(10), 3022–3024 (2005)

Surface Texture for Water Lubrication

KOJI KATO

Department of Mechanical Engineering, College of Engineering Nihon University, Koriyama, Japan

Synonyms

[Textured ceramic surfaces](#); [Water-lubricated bearings](#)

Definition

Surface texture is the three-dimensional micro geometry formed on a solid surface; it influences properties of friction and wear of solids lubricated by water.

Scientific Fundamentals

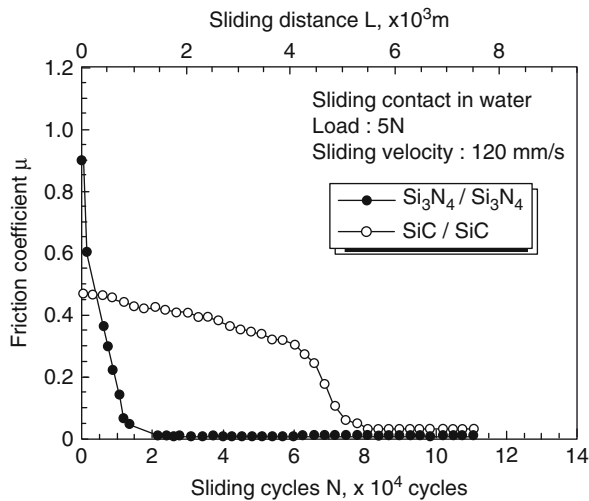
Mechanisms

Water lubrication has a friction coefficient of about 0.01 or less when contact surfaces are smooth, in the order of a nanometer. A water film, with a thickness in the tens of nanometers, is formed hydrodynamically and is not ruptured by surface asperities or surface deformation.

This condition is attained on real contact surfaces when successive wear generates smooth wear surfaces. It occurs on the contact surfaces of silicon carbide (SiC) and silicon nitride (Si_3N_4) sliding between themselves in water at room temperature. The wear mode of SiC or Si_3N_4 in water begins with mechanical wear, resulting in high friction, and transits to tribochemical wear, resulting in low friction in repeated sliding.

In tribochemical wear, silica (SiO_2) is formed on the wear surface by the reaction of “ $\text{SiC} + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + \text{CH}_4$ ” or “ $\text{Si}_3\text{N}_4 + 6\text{H}_2\text{O} \rightarrow 3\text{SiO}_2 + 4\text{NH}_3$ ” and dissolved in water by the reaction of “ $\text{SiO}_2 + 2\text{H}_2\text{O} \rightarrow \text{Si}(\text{OH})_4$.” The wear surface becomes smooth, in nanometer order, and SiO_2 staying on the surface forms silica gel ($\text{SiO}_2 \cdot n\text{H}_2\text{O}$). A friction coefficient in the range of 0.01–0.001 is generated between such surfaces by forming a water film, even when the sliding velocity is in the order of 10 cm/s under high contact pressure, in the order of 1.0 MPa.

The surface texture on a solid surface has two types: One is generally formed as a result of surface finish by machining, grinding, and/or polishing in the production process and does not have a designed pattern in the distribution of local heights at the surface in nano/micrometer scale. It is composed of waves reflecting machine dynamics for the finish and asperities reflecting machinability of the solid. The other is generally formed with micro pits and/or grooves introduced by laser etching, chemical etching, or abrasive jet etching and has a typical character in the pattern. Either type influences water lubrication and the resultant friction and wear.



Surface Texture for Water Lubrication, Fig. 1 The change of friction coefficient μ at SiC pin/SiC disk and Si_3N_4 pin/ Si_3N_4 disk caused by the repeated sliding cycles in water

A favorable surface texture exists that effectively reduces friction between the smooth wear surfaces. The friction coefficient of about 0.002 between SiC surfaces finished by polishing is reduced to about 0.0001 by introducing micro pits of a unique pattern on one contact surface.

Experimental Observations

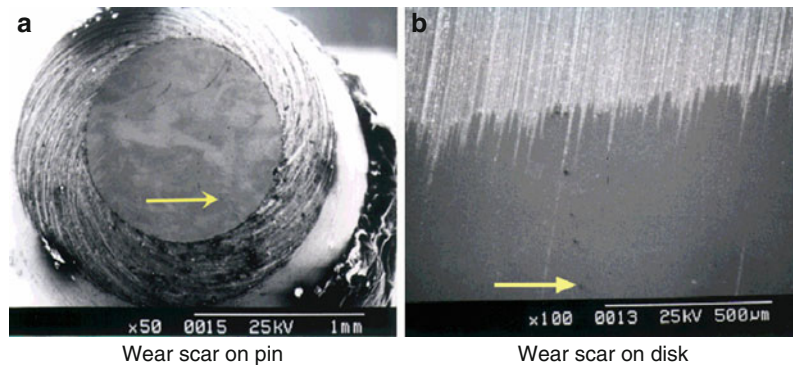
The friction coefficient μ at SiCpin/SiCdisk or Si_3N_4 pin/ Si_3N_4 disk sliding in water shows a high value in the initial running-in period of repeated friction cycles and a low value in the following steady state where μ becomes about 0.01 or less (Fig. 1).

The wear scars observed by the SEM on the SiC pin and SiC disk at the last cycle in Fig. 1 are smooth and do not show the typical fine grooves indicating the sliding direction in the images of Fig. 2a, b. Wear rate, which generates such smooth wear scars, generally stays below $10^{-7} \text{ mm}^3/\text{Nm}$.

When the sliding contact is made in water between the flat surfaces of SiC disk and cylinder that have the distribution of peak highs of asperities as shown in Fig. 3, high asperity tips contact first and experience the running-in process and steady state.

As a result, the friction torque T between flat surfaces of SiC pin and SiC cylinder changes, as shown in Fig. 4, in a step-wise loading process during repeated sliding cycles in water.

The steady state value of friction coefficient μ at the same SiC flat surfaces sliding in water changes by the change of revolution speed N for sliding and by the mean contact pressure P_m , as shown in Fig. 5, where the relationship between them changes by the change of initial

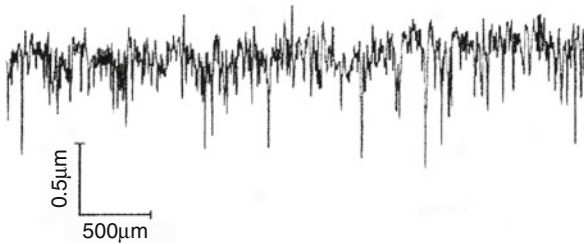


Surface Texture for Water Lubrication, Fig. 2 Smooth wear scars on the pin and disk observed by the SEM after having steady state in water. (a) Wear scar on pin, (b) Wear scar on disk. The arrows show the direction of sliding of mating surfaces

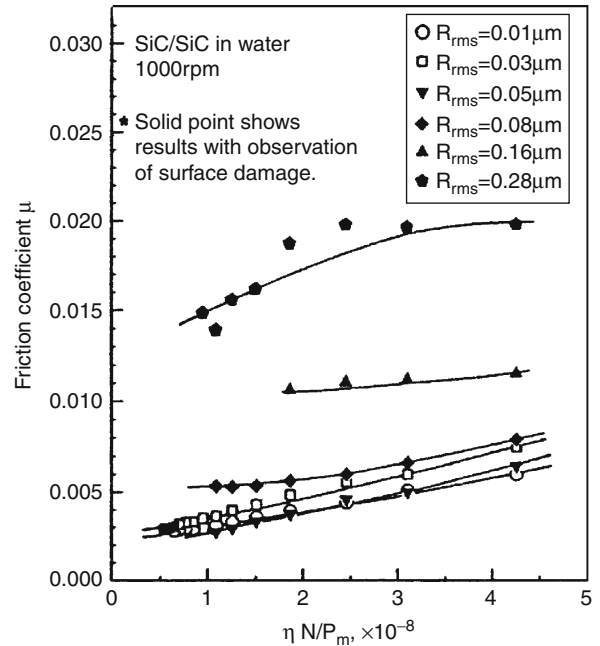
surface roughness described by the root mean square value R_{rms} .

Table 1 shows three patterns of arrangement of small and large pits fabricated on the flat surface of a SiC disk that has an average surface roughness of around $0.02\ \mu\text{m}$. These patterns are named RSP, CLP, and CLRSP in the following figures.

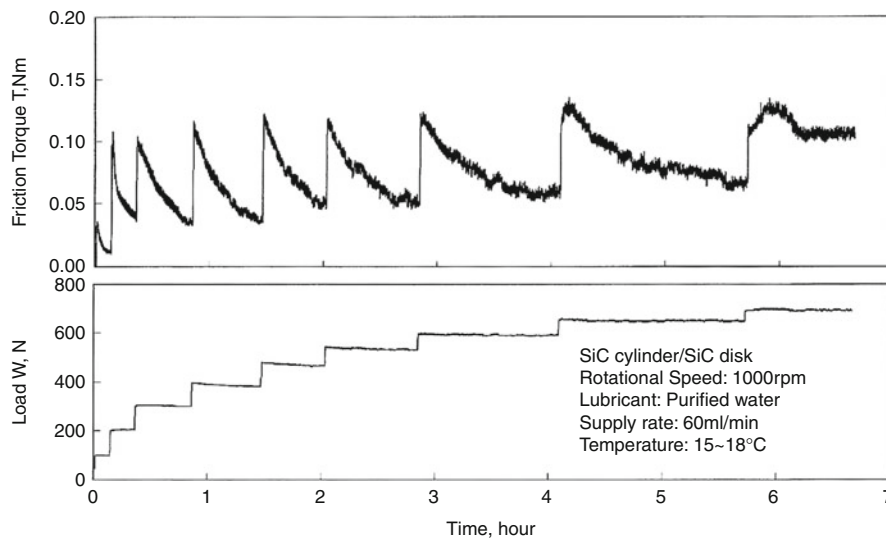
Figure 6 shows the effect of pit pattern on the friction coefficient between the flat surfaces of the disk and cylinder described in Table 1. Among the three patterns, the RSP pattern has the smallest friction coefficient of $\mu = 0.0001$, which is about one ninth of the μ value observed with the disk with no pit.



Surface Texture for Water Lubrication, Fig. 3 A typical surface roughness profile of a SiC disk finished by grinding

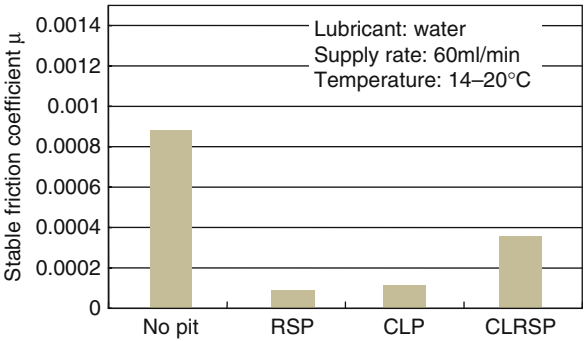
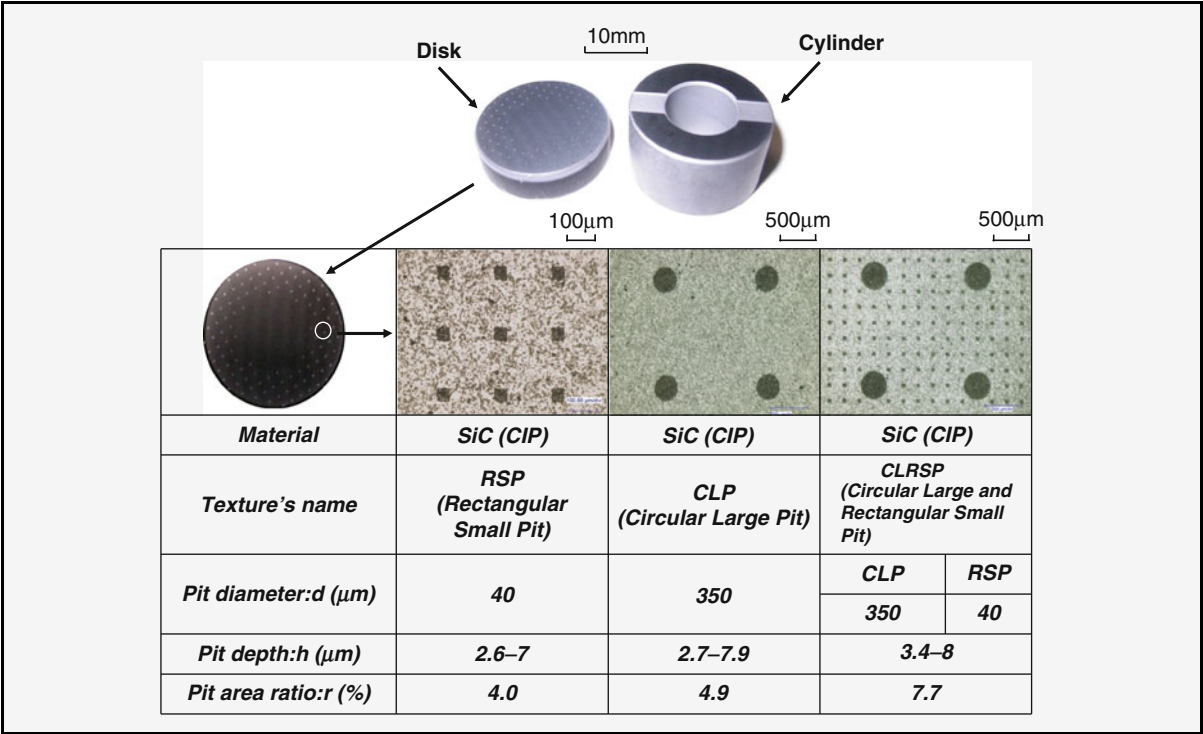


Surface Texture for Water Lubrication, Fig. 5 The relationship between the friction coefficient μ and $\eta N/P_m$ (water viscosity revolution speed/mean contact pressure), and the initial surface roughness R_{rms} on it in sliding contacts between SiC flats

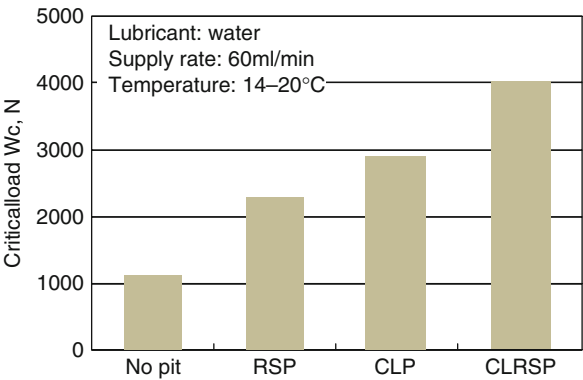


Surface Texture for Water Lubrication, Fig. 4 The change of friction torque T between flat surface of SiC pin and SiC cylinder in step-wise loading process during repeated sliding cycles in water

Surface Texture for Water Lubrication, Table 1 Three patterns of pits of RSP, CLP, and CLRSP formed on a SiC disk surface finished by grinding and polishing



Surface Texture for Water Lubrication, Fig. 6 The effect of surface texture of a SiC disk on the friction coefficient μ in the steady state in water



Surface Texture for Water Lubrication, Fig. 7 The effect of surface texture of a SiC disk on the critical load W_c for seizure in water

Figure 7 shows the effect of pit pattern on the critical load W_c for seizure in water observed with the same three patterns of pits as shown in Table 1. Among the three patterns, the CLRSP pattern has the largest critical load of $W_c = 4,000$ N, which is about four times larger than that on no pit surface.

Key Applications

Water is an effective lubricant for SiC sliding against itself at room temperature. Therefore, SiC is already in use with the journal bearings of machinery that operate in water. Good surface texture design is the key to better

performance of the bearings. However, the design standard or theory for the optimum pattern of a texture is not yet established for the general use.

Si_3N_4 is another useful material for water lubrication, as shown in Fig. 1. It is already in use with ball bearings used in water. A coating of carbon nitride (CN_x) on one side of contact surfaces of SiC or Si_3N_4 further decreases friction and wear in water.

Cross-References

► Water-Lubricated Bearings

References

- M. Chen, K. Kato, K. Adachi, Friction and wear of self-mated SiC and Si_3N_4 sliding in water. *Wear* **250**, 246 (2001)
- X. Wang, K. Adachi, K. Otuka, K. Kato, Optimization of the surface texture for silicon carbide sliding in water. *Appl. Surf. Sci.* **253**, 1282 (2006)
- H.C. Wong, N. Umehara, K. Kato, The effect of surface roughness on friction of ceramics sliding in water. *Wear* **218**, 237 (1998)
- F. Zhou, K. Adachi, K. Kato, Friction and wear property of a CN_x coatings sliding against ceramic and steel balls in water. *Diamond Relat. Mater.* **14**, 1711 (2005)

Surface Texture Generation with a Numerical Process

Q. JANE WANG¹, DONG ZHU²

¹Department of Mechanical Engineering and Center for Surface Engineering and Tribology, Northwestern University, Evanston, IL, USA

²State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing, People's Republic of China

Synonyms

Numerical surface texturing

Definition

Surface textures are formed by surface micro geometry. Virtually, textures can be numerically generated for design and analysis purpose. Surface textures thus developed can be further analyzed with a model-based simulation system to evaluate their potential tribological significance.

Scientific Fundamentals

Surface Texture Description

Textures can be made with either continuous or discontinuous features. Dimples and short grooves are typical

discontinuous textures, and grooves and waves are often seen as continuous textures. A numerically generated dimple surface is illustrated in Fig. 1 and a continuous numerical sinusoidal wavy surface is shown in Fig. 2. Figure 1 also shows the cross-sectional view of one of the numerical elliptical dimples and lubrication with the dimpled surface when it is in a relative motion with respect to its mating surface. Note that for lubrication purposes the dimple depth is usually much smaller than the lateral dimensions, and it is exaggerated in the cross-sectional view in order to show the shape clearly.

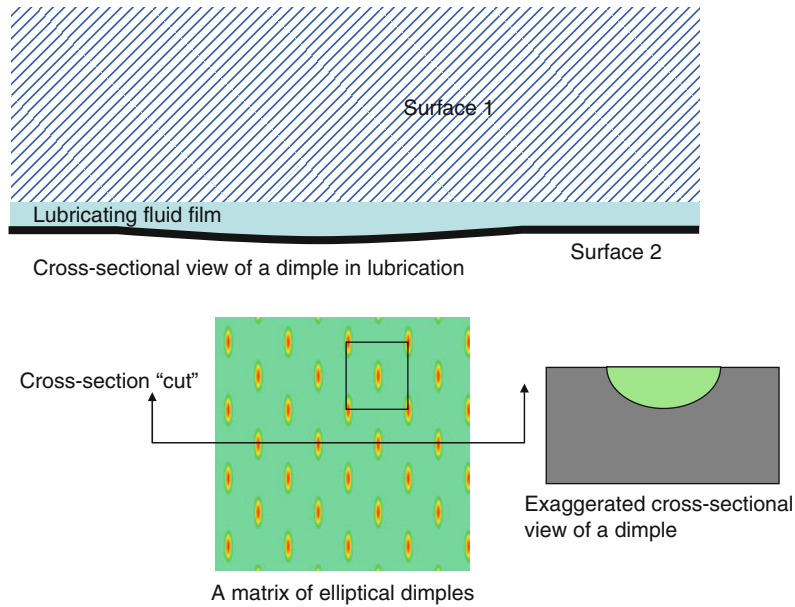
Important factors to define a texture should include the surface shape function (top shape), cross-section shape function (bottom shape), distribution function, and depth. They are independent and all should be defined in texture generation. The texture density, which can be referred to as the textured area divided by the nominal surface area, depends on the top shape and the distribution function. A texture feature may have different cross-sectional shapes and different feature distributions. Figures 3 and 4 present several bottom shapes and distribution patterns, respectively. The exact triangular shape is determined by the location of the summit, as shown in Fig. 3. Adjusting distance x to the far right, far left, and the center results in three typical triangles. W shapes are combinations of triangular shape; and they are symmetrical and advantageous with respect to motion direction reversion.

Figure 5 shows a complete description of a W-type texture. The top shape, bottom shape, and distribution functions are all given.

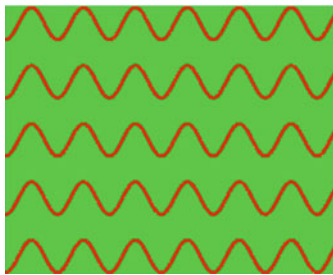
Surface Texture Numerical Generation

A textured surface can be numerically described with a unit cell of the desired geometric feature that is periodically extended to the entire surface. In most cases, therefore, surface texture generation consists of two steps: (1) numerical description of the unit cell, as indicated by the rectangular box in Fig. 1, which may be called numerical dimpling, and (2) numerical feature extension to duplicate automatically the feature following a certain pattern, which may be called numerical patterning. The textures shown in Fig. 4 were all produced from a characteristic unit cell of certain three-dimensional shapes, which were then periodically copied following certain patterns of distribution.

An elliptical dimple may be defined with its principal radii, the center span, and depth. A dimple density factor may be defined in terms of the ratio of the dimpled area to the total nominal area. The cross-sectional shape (bottom shape) can be U, R, or T shown in Fig. 3, or others.



Surface Texture Generation with a Numerical Process, Fig. 1 Numerically generated elliptical dimples. The arrows indicate the direction of the cross-sectional cut. The rectangular box indicates a unit cell



Surface Texture Generation with a Numerical Process, Fig. 2 A sinusoidal wavy texture propagated along the motion direction. Different sinusoidal textures can be obtained through changing the description function

Short grooves can be defined by width \times length \times X-direction distance \times Y-direction distance \times depth, together with the orientation angle with respect to the Y direction. The cross-sectional shape (bottom shape) can be U, R, T, or W shown in Fig. 3, or others.

Sinusoidal patterns can only be grooves. However, they can have any bottom shapes such as those shown in Fig. 3.

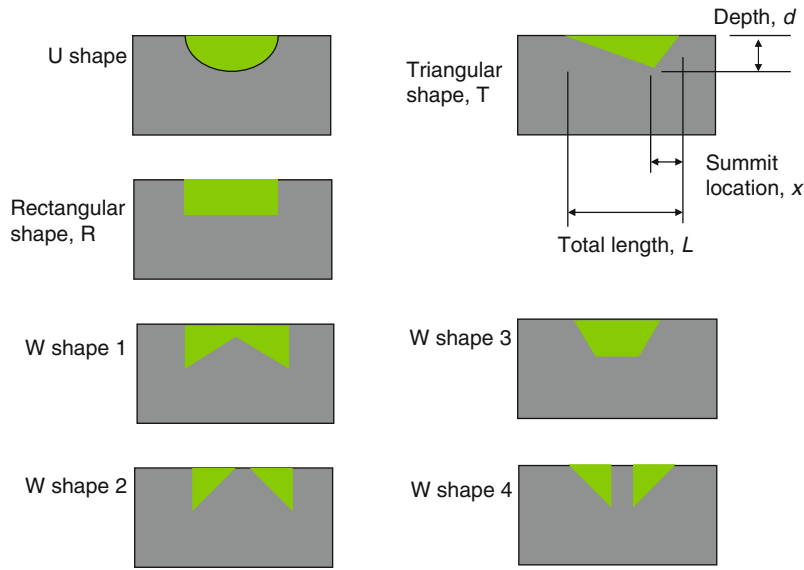
Textures with a sinusoidal pattern can be defined by amplitude \times wavelength \times groove width \times depth \times wave distance.

A key issue is to determine the dimple depth, d , which can be analytically estimated based on the theory of hydrodynamic lubrication (Pinkus and Sternlicht 1961) and the projected operating film thickness. The maximum load capacity of a slider bearing can be expressed as:

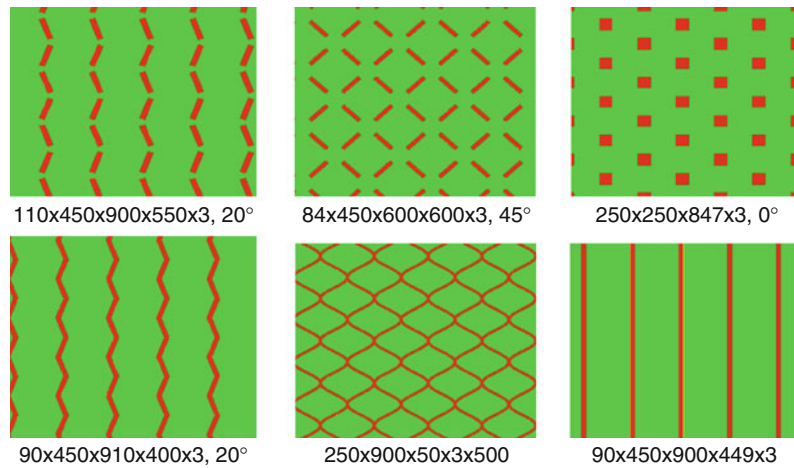
$$W_{\max} = k \frac{\eta V L B^2}{h_2^2} \quad (1)$$

In this equation, h_2 is the lubricant film thickness, V the sliding speed, L the width of the slider, and B the length of the slider. Parameter k varies for different slider bearings determined by the optimized ratio of the inlet film thickness over the outlet film thickness, h_1/h_2 . The film ratio, h_1/h_2 , may be expressed by $1 + d/H_c$ using H_c , the central film thickness, to approximate the film thickness in most of the elastohydrodynamic (EHL) region. Thus, the depth of a texture may be approximated by $d \sim H_c O(h_1/h_2 - 1)$. For example, if H_c is about 1.6μ , the optimal values of the film ratio, h_1/h_2 , for different sliders should be in the range of 1.87–2.2 (Nanbu et al. 2008).

Another way to estimate the dimple depth is by using the inlet suction mechanism. This micro-hydrodynamic lubrication mechanism was recently reported by



Surface Texture Generation with a Numerical Process, Fig. 3 Several possible bottom shapes of textures. The U, R, and T are of single geometry, while the W's are of combined geometries

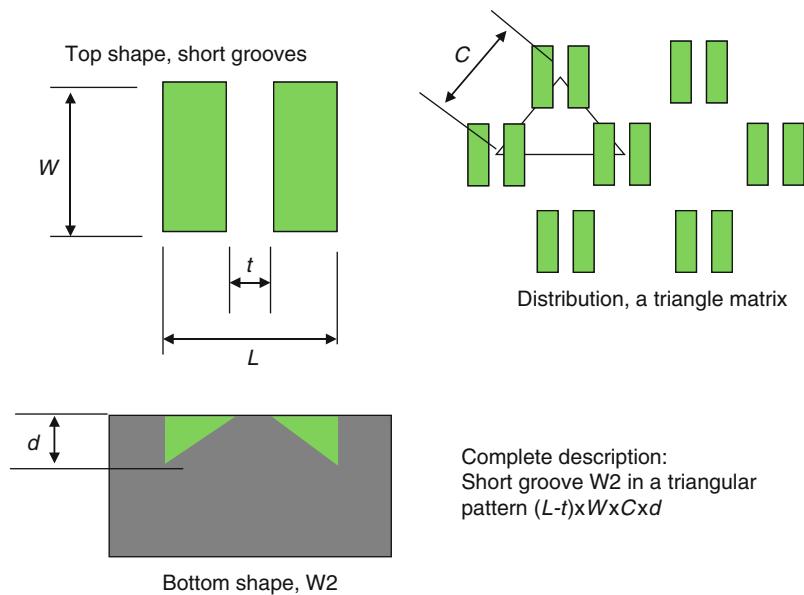


Straight grooves defined by Width x Length x X-direction distance x Y-direction distance x Depth and the orientation angle with respect to the Y direction, and sinusoidal grooves defined by Amplitude x Wavelength x Groove width x Depth x Wave distance. The unit is micron.

Surface Texture Generation with a Numerical Process, Fig. 4 Several possible distribution patterns of textures. *Top*: discontinuous textures; *Bottom*: continuous textures

Olver et al. (2006) and Fowell et al. (2007) in the study of micro-lubrication and loading lifting in textured bearings. Based on the inlet suction mechanism, Fowell et al. (2007) showed the volume flow for a parallel bearing with flat bottom dimples, or the R-type bottom shapes:

$$Q = \frac{Uh_0}{2} + \frac{Uh_0^3}{2} \frac{h_1/h_0 - 1}{(h_1/h_0)^3 (L_1 + L_2)/L + 1} \quad (2)$$



Surface Texture Generation with a Numerical Process, Fig. 5 A complete description of a W-type texture

Here, U is the sliding speed, h_0 and h_1 are the film thickness in flat and packet section, respectively. L_1 , and L_2 are length of the leading and trailing flat section, and L is the length of the packet section. The optimal pocket depth can be found by maximizing the volume flow, Q , with respect to h_1/h_0 . With $(L_1 + L_2)/L$ between 1 and 10, the ratio of (h_1/h_0) should be about 1.7 in order to have the largest volume flow rate. This result is consistent with the result described above.

However, a practical depth cannot be too shallow because the fabrication difficulty and wear must be considered. Wear and lubrication tests targeting particular applications are always useful approaches to determine depth of texture features Nanbu et al. (2008) and Wang and Zhu (2005) recommended the minimum depth of 3 micron for dimples developed for lubrication applications.

Surface Numerical Evaluation

Surface features may be considered as micro-bearings in many application cases. Features favoring lubrication cannot be designed too deep because significant pressure reduction due to dents should be avoided, otherwise the capacity of load supporting of contact may be seriously lowered. In addition, shallow features introduce less damage to the surface and subsurface material. Most machined surfaces contain shallow topographic feature. Dimpled surfaces, as an example of patterned surfaces, may be considered as integration of micro slider bearings when

it is engaged with its mating surface. Typical dimples, for example, have large size-to-depth ratios. The edges of dimples may be compared with step bearings that are proven to be the desirable geometry of slider bearings. According to the hydrodynamic lubrication theory and engineering practice with step bearings, grooved bearings and seals, turbulence caused by the edges of surface features may not be significant, and, in many cases, may be negligible. Suffice it to say that the Reynolds equation is still applicable to the lubrication of textured surfaces designed for lubrication applications.

Surface Fabrication

Micro-machining, metal-forming, electric discharge machining (EDM), vibration turning, electrochemical machining (ECM), laser or electron beam texturing, lithography, reactive etching, and mask blasting, as well as shot peening, have all been used for surface texturing. Micro-machining can generate pre-described three-dimensional (3D) surface dimples in almost any shape on any surface with ultra high-precision. Micro-forming is capable of efficiently producing high precision 3D dimples on large surfaces, but tool wear problem is a serious concern. Micro-stamping creates surface textures through micro presses of desired shape. One feature of the surfaces made by micro stamping is the compressive stresses built into the working surface. EDM is independent of workpiece hardness, but requires the workpiece to be conductive. In EDM, the forces due to pressure can result in

workpiece distortion when machining very small features. ECM can create a stress-free, undamaged surface, and has the major advantage of zero tool wear. Electron beam texturing has the advantage of higher degree of control; however, it requires vacuum and is more expensive. The laser beam process is efficient with well-controlled distribution patterns, but it is difficult to obtain accurate depth and desirable bottom shape geometry. Micro molding through lithography etching is commonly used in micro-fabrication of textured surfaces, in which an elastomer, poly (dimethylsiloxane, PDMS) is usually used to cast against a patterned substrate by lithographic techniques. Apparently, this technique works efficiently for small flat samples but requires a clean room. A lathe vibration turning approach has been developed for rotational surfaces and for the convenience in linking with the current manufacturing technologies, which can be directly applied to machine dimples on cylindrical surfaces.

Key Applications

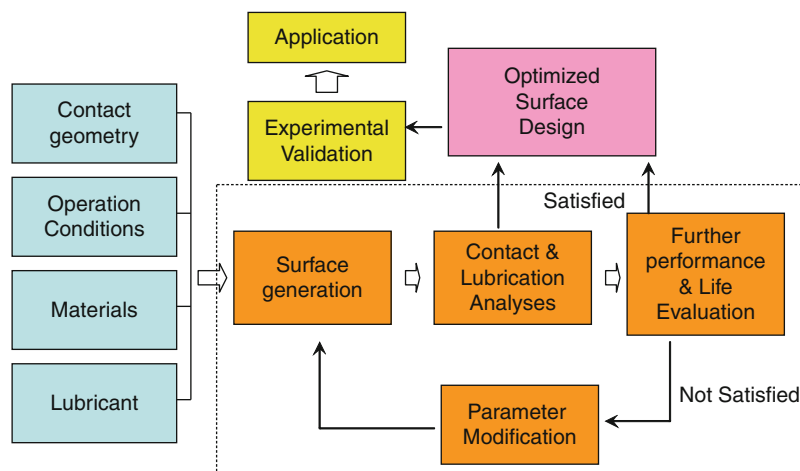
Numerical Generation of Surfaces for Texture Optimization

Surfaces with controlled micro geometry may help improve the performance of tribological interfaces, because well-designed micro features may be micro reservoirs to retain lubricant and act as micro bearings to enhance lubrication. Automotive engine pistons and cylinder liners, mechanical seals, and magnetic storage are some of the examples of texture surface applications. Developing engineered surfaces demands an in-depth understanding of the contact mechanics and mixed

lubrication of surfaces. One may anticipate that the design of surface textures should be related to the type of contact, materials in contact, and most complicatedly, the operating conditions. Optimization of a surface texture design usually requires the study of a large number of geometric parameter combinations and extensive investigations of the surface performance in an expected working environment. Experimental research of surface textures requires sample manufacturing and intensive testing, which can be expensive and time consuming even for the simplest surface patterns. Most importantly, a complete evaluation is sometimes nearly impossible because of the limitation to sample fabrication and experimental condition control. Numerical texturing can precisely produce the features of engineered surface topography and support a model-based numerical tool capable of conducting contact and mixed lubrication analyses. The trend of the performance of textured surfaces can be estimated with the model-based simulation tool, and preferred surfaces can be selected for further prototyping and experimental evaluation.

Virtual Surface Texturing

Numerical texturing and model-based simulation result in a virtual surface texturing technology (Wang and Zhu 2005), which is illustrated in Fig. 6. The following parts should be included in virtual texturing: (1) virtual surface texture generation based on application requirements; (2) contact and lubrication analyses for performance evaluation; (3) efficiency, life, and surface evolution prediction; (4) surface texture modification and optimization; and (5) necessary experimental verification. Items



Surface Texture Generation with a Numerical Process, Fig. 6 Flow chart of virtual surface texturing

embraced in the dashed box are where the numerically generated surfaces are involved, where the surface micro geometry is linked to material and lubricant properties on one hand and to the contact and lubrication performance of the interface on the other hand.

Numerical generation of surfaces is not limited to textured surfaces with precisely defined periodic features. Generation of rough surfaces with certain statistic characteristics has been intensively practiced in tribology (Patir 1978; Hu and Tonder 1992; Patir and Cheng 1979).

Cross-References

- [Laser Texturing, Characterization and Related Effects](#)
- [Surface Texture for Magnetic Recording](#)
- [Surface Texture for Water Lubrication](#)
- [Surface Texturing by Vibro Machining](#)

References

- M. Fowell, A.V. Olver, A.D. Gosman, H.A. Spikes, I.G. Pegg, Entrainment and inlet suction: two mechanisms of hydrodynamic lubrication in textured bearing. *ASME J. Tribol.* **129**(2), 336–347 (2007)
- Y.Z. Hu, K. Tonder, Simulation of 3-D random rough surface by 2-D digital filter and Fourier analysis. *Int J Mach Tools Manuf* **32**(1–2), 83–90 (1992)
- T. Nanbu, N. Ren, Y. Yasuda, D. Zhu, Q. Wang, Micro textures in concentrated conformal-contact lubrication: effects of texture bottom shape and surface relative motion. *Tribol. Lett.* **29**, 241–252 (2008)
- A.V. Olver, M.T. Fowell, H.A. Spikes, I.G. Pegg, 'Inlet suction', a load support mechanism in non-convergent, pocketed, hydrodynamic bearings. *Proc. IMechE Part J. J. Eng. Tribol.* **220**(2), 105–108 (2006)
- N. Patir, Effect of surface roughness on partial film lubrication using an average flow model based on numerical simulation, Ph.D. thesis, Northwestern, (1978)
- N. Patir, H.S. Cheng, Application of average flow model to lubrication between rough sliding surfaces. *ASME J. Lubr. Technol.* **101**(1979), 220–230 (1979)
- O. Pinkus, B. Sternlicht, *Theory of Hydrodynamic Lubrication* (McGraw-Hill, New York, 1961)
- Q. Wang, D. Zhu, Virtual texturing: modeling the performance of lubricated contacts of engineered surfaces. *ASME J. Tribol.* **127**, 722–728 (2005)

Surface Texturing by Laser Ablation

GABRIEL DUMITRU

BCI Group, Grenchen, Switzerland

Synonyms

[Laser structuring](#); [Patterning of tribological surfaces](#)

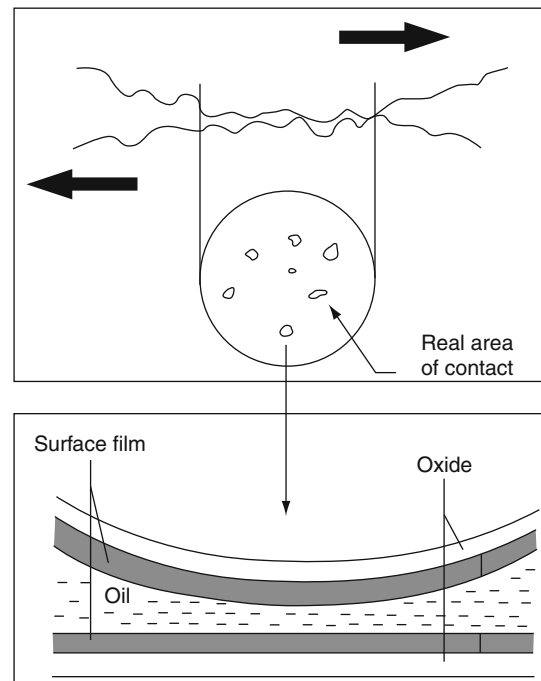
Definition

The ablative laser structuring of tribological surfaces denotes the generation of microscopic morphology or structure modifications of tribologically stressed surfaces by laser material removal.

Scientific Fundamentals

Friction, Wear, Surface Texturing

Due to the intrinsic surface roughness, the contact in macroscopic tribological systems occurs actually on a microscopic scale: surface protuberances, on which large mechanical stresses are concentrated and whose temperatures can increase rapidly (Fig. 1). Many systems operate under high loads and/or slow speeds that do not allow building up an effective lubricant film, fully separating the surfaces. Furthermore, due to inherent phenomena or functioning conditions (e.g., stress cycles, frequent start-ups, accelerate aging, repeated heating, and cooling), it cannot be avoided that some protuberances loosen themselves from the surfaces, becoming debris particles with an abrasive effect. As long as these particles remain small and the surfaces are separated by the lubricant film, the potential damages are limited; the eventually occurring system breakdown is related to a discontinuous



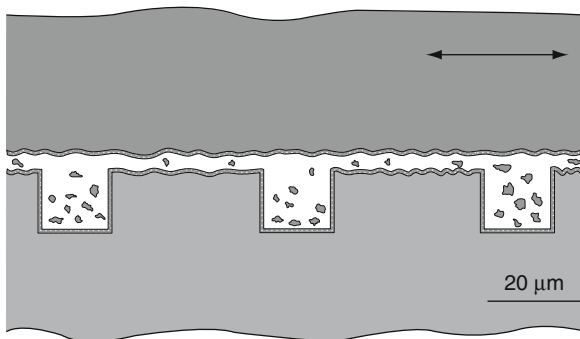
Surface Texturing by Laser Ablation, Fig. 1 Surface contact in tribological systems

lubricant film or to large wear particles, generally arising from surface fatigue fractures.

The wear in tribological systems can be reduced (and their breakdown thus delayed) by minimizing the formation of debris particles or by removing them from the system. The former can be achieved by using hard materials, protective coatings, or additives that prevent the occurrence of unwanted chemical reactions. This can be accomplished by circulating and filtering the lubricant, by using a softer surface as second gliding partner, or by appropriate geometrical modifications of the contact surfaces. In this last case, an array of uniformly distributed pores is produced (e.g., by laser ablation) on a gliding surface (Fig. 2). Besides trapping the wear particles, these arrays may enhance further beneficial mechanisms: they represent lubricant reservoirs and can prolong the presence of the protective film in starved lubrication conditions and, depending on contact pressure, sliding velocity, and lubricant viscosity, they can enhance the hydrodynamic and hydrostatic lubrication of many tribological components.

For a holistic approach, which allows also a better understanding of the aforementioned beneficial aspects, the wear-related considerations must be completed with other physical and chemical properties of the surface material and the lubricant. For instance, the viscosity of the lubricant determines the film thickness, and the roughness of the contacting surfaces has a major impact on their friction and wear behavior (Fig. 3); their ratio λ gives an indication of the effectiveness of the fluid lubrication mechanisms. Low λ values correspond to the previously considered situation, with severe contact conditions and higher wear.

Arrays of uniformly distributed pores were reported to expand the limits of the hydrodynamic lubrication regime for both high- and low-viscosity oil lubricants (λ shift)



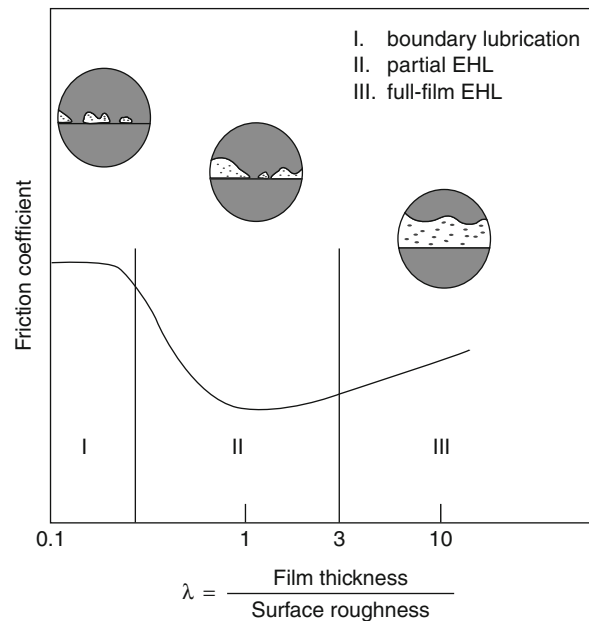
Surface Texturing by Laser Ablation, Fig. 2 Pores as debris particle traps and lubricant reservoirs

and were also observed to reduce friction in oil-lubricated tribological components operating under a boundary lubrication regime (low λ values).

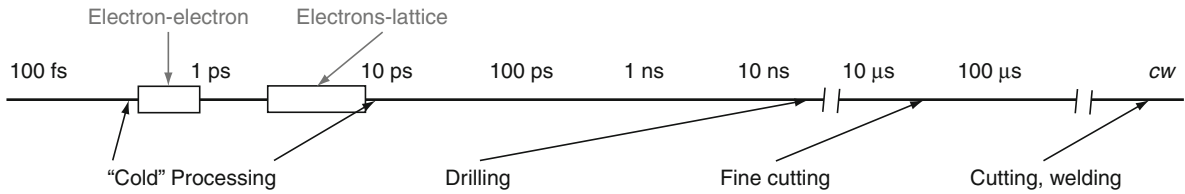
Ablative Laser Texturing

Structures on hard surfaces (e.g., arrays of dots, lines, or grooves) can be fabricated using different manufacturing processes. Surface texturing by vibrorolling consists of producing shallow grooves by plastic deformation using a hard indenter on metallic parts. Reactive ion etching was also employed to texture sliding SiC surfaces, and LIGA (Lithography, Electroplating, and Molding) techniques are very efficient in producing structures with dimensions below 100 nm on flat substrates. Electro-discharge machining can be also used to texture metals, as wire-electrodes in the range of 10–100 μm are available.

Some of these procedures are suitable for mass production; however, they require large, expensive facilities, can show process-related draw-backs (e.g., processing of non-planar, hard, and or non-conductive surfaces), and are not always suitable for flexible production. In such situations, laser texturing may be the technique of choice. By controlling the incident energy density and its distribution on the treated surface, the laser can produce arbitrary patterns on hardened steels, ceramics, polymers, and crystalline structures. This potential, combined with recent improvements of available laser sources, leads to



Surface Texturing by Laser Ablation, Fig. 3 The Stribeck curve



Surface Texturing by Laser Ablation, Fig. 4 Thermalization times and some laser processes

a continuous growth in the use of lasers as surface texturing tools.

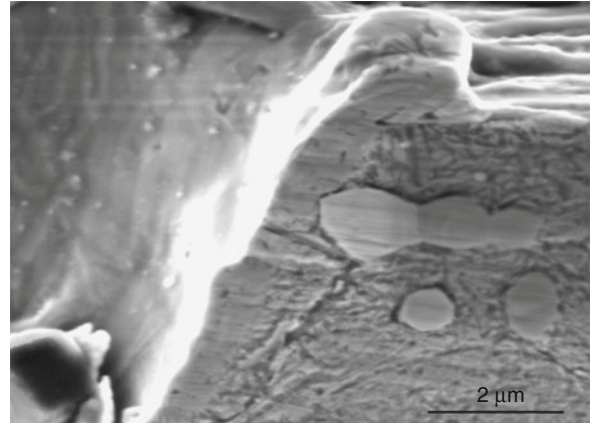
In discussing the particularities of the surface texturing by laser ablation, some aspects of laser-matter interactions should be considered. A laser beam that is focused on a surface generates an intense electric field, localized under the irradiated surface. Electrons are accelerated by this field and gain kinetic energy; due to their mobility, they collide with lattice atoms and transfer this energy to them. The vibration energy of lattice atoms is macroscopically mirrored in material heating and in phase changes. This photons-electrons-phonons energy transfer chain (Fig. 4) takes place in about 1 ps in metals and takes slightly longer in ceramics.

For nanosecond pulses (e.g., excimer or Q-switched lasers), the energy transfer occurs during the pulse, under thermal equilibrium conditions, and material removal takes place mostly through melting and vaporization. In describing these laser-induced thermal phenomena, the thermal diffusion length l is a very useful parameter for a first approximation and it depends on the pulse duration τ_p , on the thermal conductivity k , on the mass density ρ , and on the heat capacity c of the considered material:

$$l = \sqrt{\frac{k}{\rho c} \tau_p}$$

For example, in the case of steel being textured with 100 ns laser pulses, the calculated thermal diffusion length is 0.8 μm . This indicates that the extension of the thermal affected zone surrounding a 10- μm pore produced using Q-switched laser pulses can be smaller than 10% of its pore diameter, as can be seen in Fig. 5.

For laser pulses shorter than the thermalization time (femtoseconds and a few picoseconds), the energy transfer occurs first in a superficial layer under non-equilibrium conditions. Hot electrons are generated, whereas lattice atoms still have undisturbed energies. In this phase, the ablation occurs through different non-thermal mechanisms (e.g., induced local stresses, Coulomb explosion, material breakdown). Following the laser pulse, particularly at high incident fluences, a part of the initial pulse

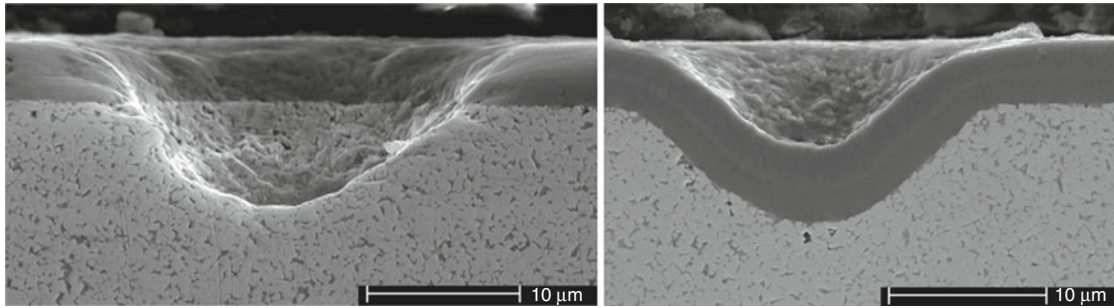


Surface Texturing by Laser Ablation, Fig. 5 Detail from the wall of a pore (steel) produced with nanosecond pulses

energy might still be stored in the hot electrons and it is transferred to the lattice after the pulse, leading to in heat-flow-driven processes.

The extension of the thermal-affected zone, the pulse duration, and the incident energy density of the texturing laser beam are important elements in defining the texturing strategy for coated surfaces: (1) direct processing, where a previously coated part is laser textured, or (2) indirect processing, where the part is first laser textured and then coated (Fig. 6). These strategies have different technological requirements, their optimization paths are dissimilar, and they usually involve different laser sources.

For instance, the highest priority in direct texturing is laser machining without any collateral thermal damage. The wavelength and the pulse duration of the processing laser must match the thermal and optical properties of the processed coating. In this case, small ($< 1 \text{ J/cm}^2$) incident energy densities are used, ultraviolet (UV) or femtosecond (fs) lasers have certain advantages, and material removal rates are rather low. On the other hand, indirect texturing is optimized for rapid and versatile material removal from metallic substrates (e.g., steel, hardmetal) and it can be performed efficiently with infrared (IR) nanosecond (ns)



Surface Texturing by Laser Ablation, Fig. 6 Pores from direct (left) and indirect laser textured surface, TiCN on hardmetal

lasers. Large ($20\text{--}100\text{ J/cm}^2$) laser energy densities are here involved, whereas high ablation rates and patterning flexibility are required. A soft mechanical polishing is carried out after laser processing, before the coating procedure.

Key Applications

Laboratory Tests

Arrays of pores, grooves, and other periodical laser-induced patterns are used in connection with steels, hardmetal, sapphire, silicon, SiC, or Ti-Al alloys, as well as in relation with a large variety of coatings: DLC, TiN, TiCN, ZrO_2 , MoS_2 , and Ni-P plated substrates (Dumitru 2007). The behavior of such textured surfaces is primarily tested using lab tribometers that simulate (more or less accurately) the real functioning conditions. The measured friction coefficients and test durations give direct indications on improvements and the analyses of the tested surfaces (e.g., SEM, EDX) provide valuable data to understand the occurring processes. Many of these tribometers rely on the ball-on-flat set-up, but other application-related reciprocating executions are also reported: block-on-ring, cylinder-on-disc, and disc-against-disc for rotation, as well as flat-on-flat for translation.

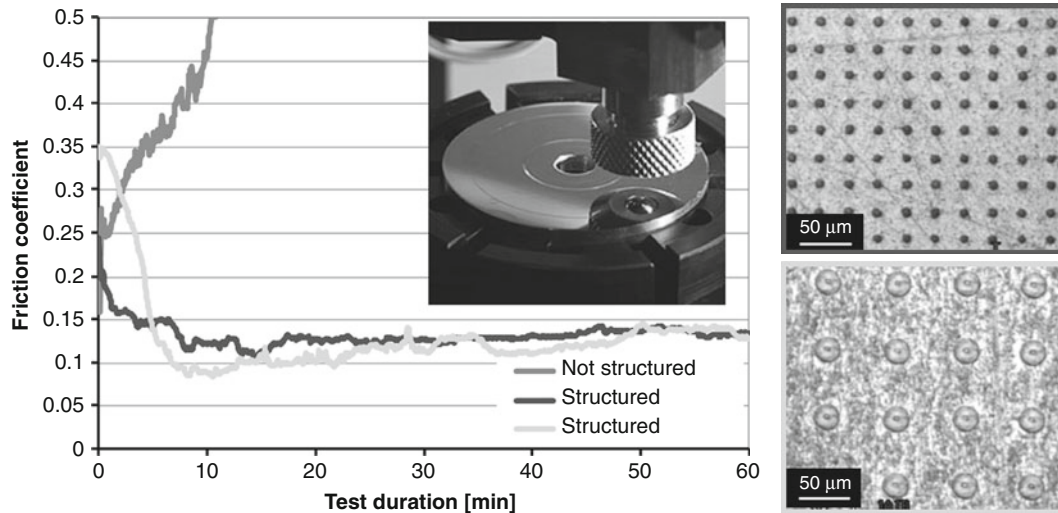
In a standard ball-on-disk apparatus a ball is loaded against a flat disk specimen surface, the direction of loading is parallel to the axis of rotation, and the pin motion describes a circular wear path. In the presented case here (tests on laser textured steel surfaces), the ball was made of hardmetal with a diameter of 6 mm and all laser-microtextured steel disks were 20 mm in diameter and 5 mm in thickness. The following conditions were used for the tribotests (Dumitru et al. 2000): sliding speed of 1 cm/s, normal force of 30 N, temperature of $22\pm1^\circ\text{C}$, ambient air with 50% relative humidity. The lubricant tested was a commercially available mineral oil with dynamic viscosity of $\eta_{40} = 96\text{ cP}$. Lubrication of the steel

disks was accomplished through a spray technique providing a lubricant volume of $0.2\text{ }\mu\text{l}$ on each sample. The tribotest was automatically stopped when the friction coefficient reached a limit value $\mu_{\text{lim}} = 0.5$ (steel-steel contact) and Fig. 7 depicts the evolution of the friction coefficient during the tribotest for two different patterns.

In a linear-oscillation tribometer, the investigated surface is flat and fixed, and a hard probe body, which is placed under a controllable load, oscillates on it with given amplitude and speed (Dumitru et al. 2005). The friction coefficient between the probe body and the tested surface is continuously monitored and measurements can be stopped either after a predefined time or if the friction coefficient exceeds a certain limit. In this case (tests on laser textured and TiN coated hardmetal parts), the cylinder was sliding with a frequency of 10 Hz and a stroke of 2 mm, and the sample was inserted in a lubricant (additivated mineral oil) containing recipient. The probe body was made of low-carbon steel and the oil covered fully the sample surface. During the first 30 s the load was increased linearly to raise the contact pressure from 20 to 250 N/mm^2 and this value was kept constant – a friction coefficient $\mu > 0.45$ or a testing time of 30 min – until the end of the test. After tribotests, EDX analyses were carried out and they revealed severe wear on the unpatterned surface and only slight wear features on the textured one. The inset in Fig. 8 depicts a SEM image from the wear track and the spatial distribution of the chemical elements Ti, N, and Fe determined by EDX mapping.

Field Test

Laboratory tests indicate the potential of the laser textures and give structure optimization trends, but the whole extent of the beneficial effects of the surface texturing is given only in production conditions. The outcome of such field test, centered on a cold metal forming tool, is presented here.



Surface Texturing by Laser Ablation, Fig. 7 Test apparatus (*inset*), tested patterns (*right*), and tribotest results

For a metal forming production step, a tool with a certain radius of curvature was designed and fabricated of hardmetal. A TiCN coating was foreseen for it. Starting from the anticipated contact pressures, a tribological surface structure was considered, implemented by laser ablation on flat surfaces and then tested through linear-oscillation experiments on flats. An optimum of the structure was determined and afterwards the most stressed tool sections, calculated by FE simulations, were laser structured. A set-up based on a 1,064 nm laser (60–80 ns, TEM₀₀, 50–100 μJ), allowing the texturing of both flat and curved surfaces was used. Roughly 200,000 dimples were produced. After the laser treatment, the tool was gently polished (Fig. 9) and then coated with a TiCN protective layer.

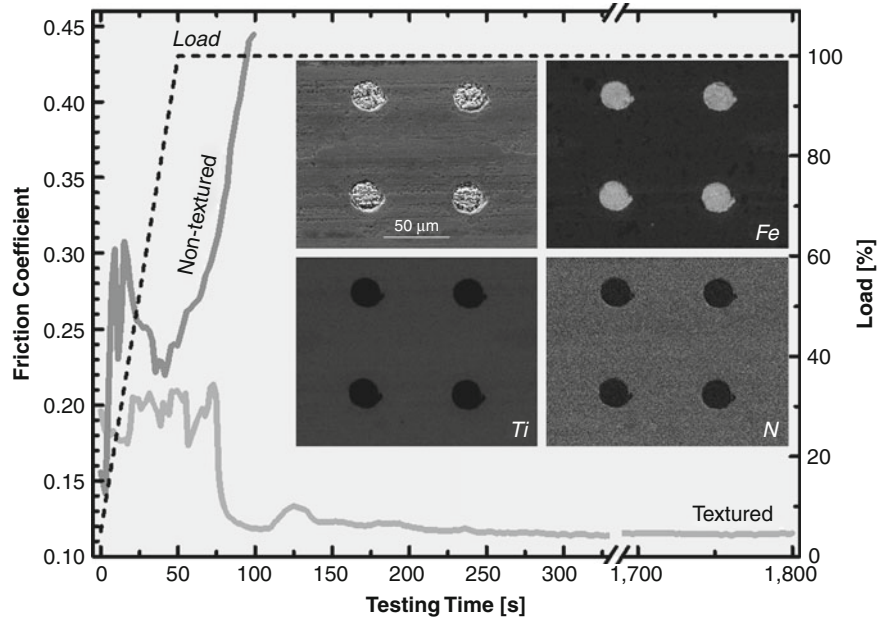
The textured tool was subsequently used in production (sheet steel) and its service life was compared with the average service life of unstructured tools T_{ref} . During operational stops, the tool was analyzed using SEM and EDX techniques. After $26 \cdot T_{ref}$, some layer delaminations were observed both in the textured and in the untextured sections. No correlation between these coating defects and the laser-produced pores was noticed. After $68 \cdot T_{ref}$, the wear was more severe, both local delaminations and ploughing traces were observed. These damages were again randomly distributed, but the pores were partially filled with debris (Fig. 10), which was determined to be Fe from the processed sheet steel. The tool had a service life of $99 \cdot T_{ref}$.

Various Applications

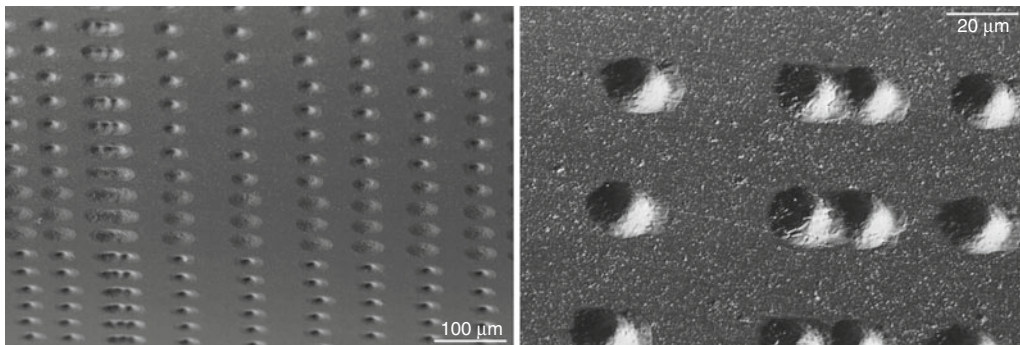
As previously mentioned, the field of laser surface texturing for tribological purposes underwent exponential growth in the recent years and various concrete industrial applications (from crude oil pumps to hard-discs) are available. Concise presentations of several uses of laser surface texturing are given below. Briefly, these applications bear on arrays of conveniently distributed laser-induced pores, which can play different roles: (1) they can promote the occurrence of hydrostatic or hydrodynamic lubrication conditions; (2) they can serve as lubricant reservoirs, capable of feeding the lubricant directly into the contact zone; and (3) they can trap wear particles and to remove them from the tribocontact area.

The technique of laser ablation was employed to create groove-type patterns of different dimensions on sapphire flats and a thin film of a liquid lubricant was deposited on the patterned surfaces (Blatter et al. 1998). Comparative pin-on-disk tests revealed a pronounced influence on the structures on the tribological characteristics of the flats; the presence of grooves drastically extended the sliding life. If the grooves were too large, the sliding life was limited by excessive wear of the pin, while with sufficiently fine grooves, a low friction steady-state regime approached, after an extended run-in period.

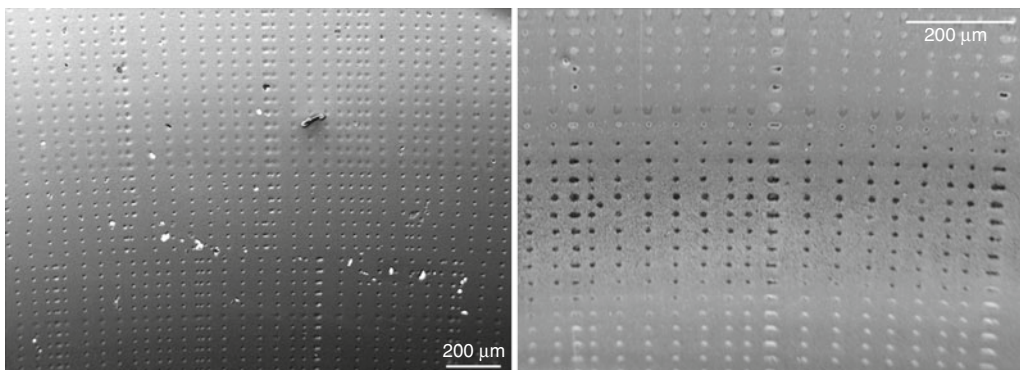
Laser texturing techniques were also used for hard-disc drives (Chilamakuri and Bhushan 1999). These discs have a landing area, where controlled roughness is created by laser texturing, and a smooth data area, where high-



Surface Texturing by Laser Ablation, Fig. 8 Measured friction curves and results of EDX analyses in wear track (*inset*)



Surface Texturing by Laser Ablation, Fig. 9 Surface of a laser-textured metal forming tool



Surface Texturing by Laser Ablation, Fig. 10 Tool surface after $26 \cdot T_{ref}$ (*l*) and after $68 \cdot T_{ref}$ (*r*)

density data is written. A laser beam is used to burn arrays of bumps in the landing area, which minimize the wear and friction on the hard drive that is due to repeated head landing on the media, when the hard drive is turned off. If the drive slows down or abruptly stops and the head lands on the smooth data area, the head scratches the media by spinning up and causes debris, which can damage other areas of the disk.

Surface structuring of SiC parts in water lubrication was also carried out by laser ablation (Wang et al. 2001). Pores with a diameter of 100–200 μm and depths of about 8–10 μm were produced by laser ablation with a CO₂ laser. Different structures were tested using a disc-on-cylinder tribometer and compared with results from the untextured specimen. An optimum value of pore area ratio was found (2.8%), giving the largest value of critical load, whereas the critical load itself was larger than that of the untextured surface by 20%. Unwanted heat-induced effects in the pores surrounding zones (with a large heat-affected zone, hardness decreased) were also observed, but these can be eliminated by using another laser source.

Extensive theoretical and experimental work was performed on laser surface texturing of mechanical seals (Etsion 2005), where the texturing was demonstrated to expand the contact parameters in terms of load and speed for hydrodynamic lubrication, as indicated by friction transitions on the Stribeck curve (Kovalchenko et al. 2005). The beneficial effects of laser surface texturing are more pronounced at higher speeds and loads and with higher viscosity oil.

The laser honing process combines laser structuring with conventional honing to produce a microstructure in cylinder walls for both high bearing capacity and excellent gliding properties. The microstructure comprises pockets (either spiral or cup structures) that are ablated into the cylinder wall by a laser beam and, after laser machining, the cylinders are finish honed to remove the melt rims on each side of the groove.

Excimer laser radiation (a mask imaging technique) was used to apply microstructures to hard-coated cold forging punches (Popp and Engel 2006). The structured punches, several coated with a MoS₂ layer, were tested in a press shop under conditions of industrial mass production. Service time was increased up to 300%.

With recent progress in novel laser sources, delivering ultrashort pulses, the direct texturing of thin hard layers becomes a more serious alternative. For instance, DLC films were nanostructured by fs-laser ablation and afterwards coated with a MoS₂ layer (Vestentoft et al. 2005), which greatly improved the frictional properties of the

system, where the smallest measured value was $\mu = 0.02$ against a hardmetal ball.

The laser-related improvements also pave the way to smaller surface structures, of great interest in microtribology. For the fabrication of such periodic nanostructures, interference techniques are ideally suited. In such schemes, the incoming laser beam is split into partial beams that are then recombined on the sample surface. The resulting interference pattern is converted into the desired surface texture (Yasumaru et al. 2008). The main drawbacks of these methods are related to the limited process field of each exposure, complicating the fabrication of large grating areas.

Solid lubricant systems can be also potentiated by ablative surface structuring (Voevodin and Zabinski 2006). These mechanisms include “on demand” solid lubricant supply from reservoirs in hard wear protective coatings. These micro-reservoirs, arrays of pores, were machined by a focused UV laser beam on the surface of hard TiCN coatings. Solid lubricants based on MoS₂ and graphite were then applied to such laser-textured surfaces and sliding friction tests were performed against steel balls in humid air and dry nitrogen environments. The service life of the solid lubricants on dimpled surfaces was an order of magnitude longer than on the unmodified TiCN coating surface.

Cross-References

- [Electron Beam Surface Technologies](#)
- [Gear Surface Treatment](#)
- [Laser Peening and Wear](#)
- [Surface Texture for Water Lubrication](#)

References

- A. Blatter, M. Maillat, S.M. Pimenov, G.A. Shafeyev, A.V. Simakin, Lubricated friction of laser micro-patterned sapphire flats. *Tribol. Lett* **4**, 237–241 (1998)
- S. Chilamakuri, B. Bhushan, Contact analysis of laser textured disks in magnetic head–disk interface. *Wear* **230**, 11–23 (1999)
- G. Dumitru, Laser processing of tribological DLC films: an overview, in *Tribology of Diamond-like Carbon Films*, ed. by C. Donnet, A. Erdemir (Springer, New York, 2007), pp. 571–590
- G. Dumitru, V. Romano, H.P. Weber, H. Haefke, Y. Gerbig, E. Pflüger, Laser microstructuring of steel surfaces for tribological applications. *Appl. Phys. A* **70**, 485–487 (2000)
- G. Dumitru, V. Romano, Y. Gerbig, H.P. Weber, H. Haefke, Femtosecond laser processing of nitride-based thin films to improve their tribological performance. *Appl. Phys. A* **80**, 283–287 (2005)
- I. Etsion, State of the art in laser surface texturing. *J. Tribol.* **127**, 248–253 (2005)
- A. Kovalchenko, O. Ajayi, A. Erdemir, G. Fenske, I. Etsion, The effect of laser surface texturing on transitions in lubrication regimes during unidirectional sliding contact. *Tribol. Int.* **38**, 219–225 (2005)

- U. Popp, U. Engel, Microtexturing of cold-forging tools – influence on tool life. *J. Eng. Manuf.* **220**, 27–33 (2006)
- K. Vestentoft, J.A. Olesen, B.H. Christensen, P. Balling, Nanostructuring of surfaces by ultra-short laser pulses. *Appl. Phys A* **80**, 493–496 (2005)
- A.A. Voevodin, J.S. Zabinski, Laser surface texturing for adaptive solid lubrication. *Wear* **261**, 1285–1292 (2006)
- X. Wang, K. Kato, K. Adachi, K. Aizawa, The effect of laser texturing of SiC surface on the critical load for the transition of water lubrication mode from hydrodynamic to mixed. *Tribol. Int.* **34**, 703–711 (2001)
- N. Yasumaru, K. Miyazaki, J. Kiuchi, Control of tribological properties of diamond-like carbon films with femtosecond-laser-induced nanostructuring. *Appl. Surf. Sci.* **254**, 2364–2368 (2008)

Surface Texturing by Vibro Machining

Q. JANE WANG¹, AARON GRECO², KORNEL EHMANN²

¹Department of Mechanical Engineering and Center for Surface Engineering and Tribology, Northwestern University, Evanston, IL, USA

²Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA

Synonyms

VMT – Vibro-mechanical texturing

Definition

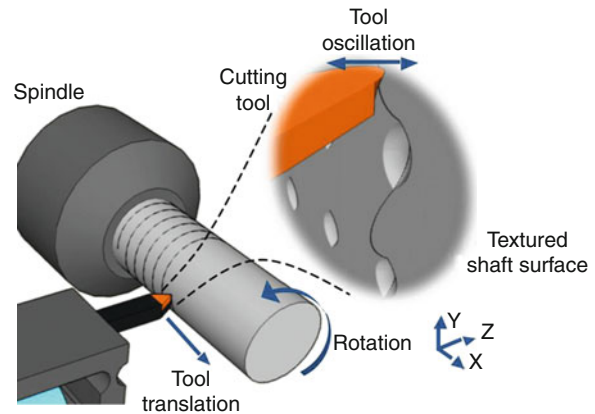
Surface Texturing by Vibro Machining, or Vibro-Mechanical Texturing (VMT), is an efficient machining method used to generate micro-texture features on the surface of a component in a tribological interface. Surface texturing is a treatment used for improving the tribological performance of a lubricated tribological interface system by enhancing the hydrodynamic action that keeps surfaces separated.

Scientific Fundamentals

Surface Topography Generation via VMT

The Vibro-Mechanical Texturing (VMT) method is based on a standard, single point turning process, with the addition of an advanced tool positioning system that allows for controlled tool vibration, illustrated in Fig. 1.

In a standard turning process, the workpiece rotates while the cutting tool engages the workpiece surface causing shear removal of surface material. The amount of material removed depends on the depth-of-cut and the tool tip geometry (e.g., nose radius, r). In the traditional turning process, there are two degrees of controlled motion: (1) the workpiece rotation, and (2) the translation of the tool across the length of the workpiece.



Surface Texturing by Vibro Machining, Fig. 1 Illustration of the VMT machining process

The VMT process introduces a third order of controlled motion, termed the “tertiary motion,” which is imposed along the axis of tool engagement, Fig. 1. This third-order tool path allows for the generation of dispersed surface features/textures (i.e., dimples) (Greco et al. 2009).

Surface Topography Calculations

In order to machine dimples on a workpiece surface, the cutting tool should oscillate at a certain frequency, f , and amplitude, G . Various waveforms can be utilized for different dimple shapes (i.e., sine, square, triangular). Figure 2 illustrates a sine wave tool path over the workpiece surface with the respective texturing parameters and resulting dimple dimensions. The neutral position refers to the position that the tool tip resides when the actuator is not energized. Δ corresponds to the separation between the neutral position to the workpiece surface.

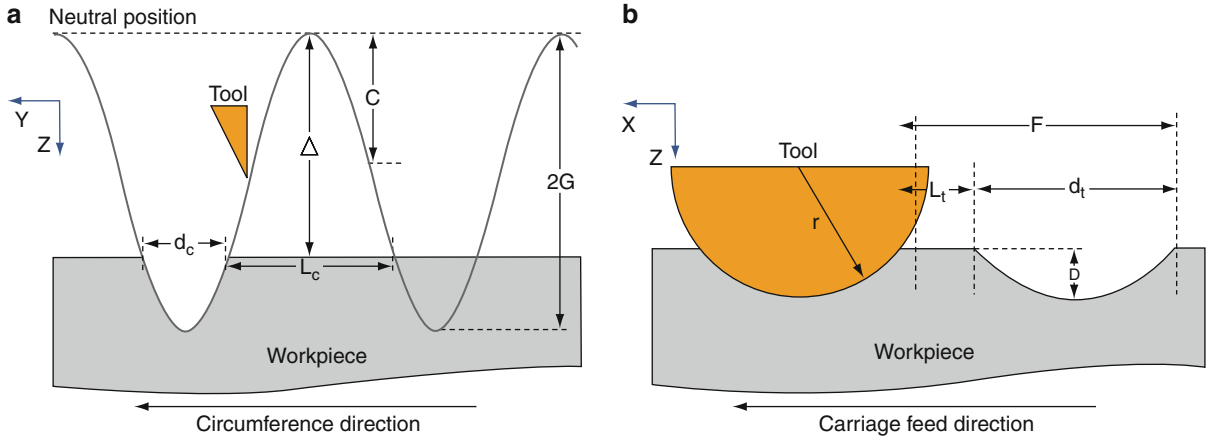
The tool path along the workpiece surface is determined from the equations describing the tool oscillation, Z , with respect to the rotational position of the workpiece, θ , in addition to the lateral translation of the tool with respect to time, $X(t)$. A sine-wave tool oscillation is considered in the following equations:

$$\text{Tool path : } Z(\theta) = G \sin\left(\frac{f}{2\pi N}(\theta)\right) + c \quad (1)$$

$$\text{Tool carriage position : } X(t) = F \times N \times t \quad (2)$$

where N is the rotational speed of the workpiece, c is the offset of the tool oscillation from the neutral position, and F is the feed rate of the tool carriage (in terms of distance/revolution).

The dimensions of the dimple pattern are calculated from the tool path (1) and (2). The main parameters that



Surface Texturing by Vibro Machining, Fig. 2 Tool path relative to workpiece surface (a) tool oscillation (b) tool carriage translation

describe a dimple pattern: axial width d_p , axial spacing L_p , circumferential width d_c , circumferential spacing L_c , and depth D are expressed as follows:

$$\text{Axial diameter: } d_t = 2\sqrt{r^2 - [r - (G + c - \Delta)]^2} \quad (3)$$

$$\text{Axial spacing: } L_t = F - d_t \quad (4)$$

$$\text{Circumferential diameter: } d_c = \frac{2\pi RN}{f} \left(1 - 2\arccos\left(\frac{\Delta - c}{G}\right)\right) \quad (5)$$

$$\text{Circumferential spacing: } L_c = \frac{4\pi RN}{f} \arccos\left(\frac{\Delta - c}{G}\right) \quad (6)$$

$$\text{Depth: } D = G + c - \Delta \quad (7)$$

Note that in these calculations, deformation restoration is not considered.

Tool Actuation

The essential component that enables the VMT machining process is the actuation system that precisely positions the tool along the axis of engagement. This system is referred to as a Micro-Positioning Stage (MPS). MPS systems are used in other similar machining processes in which the devices are referred to as Fast Tool Servos (FTS) (Trumper and Lu 2007). The earliest applications of this type of machining system was used in the intaglio, or gravure printing process (Montesanti 2005). Presently, more advanced systems are used for precision machining of shafts (Kim and Kim 2003) or manufacturing of microlens arrays (Noh et al. 2008). FTS is a term that identifies a broader family of tool actuators with a wide range of stroke amplitudes, $1\text{--}10^4 \mu\text{m}$, and bandwidth, $10\text{--}10^4 \text{ Hz}$.

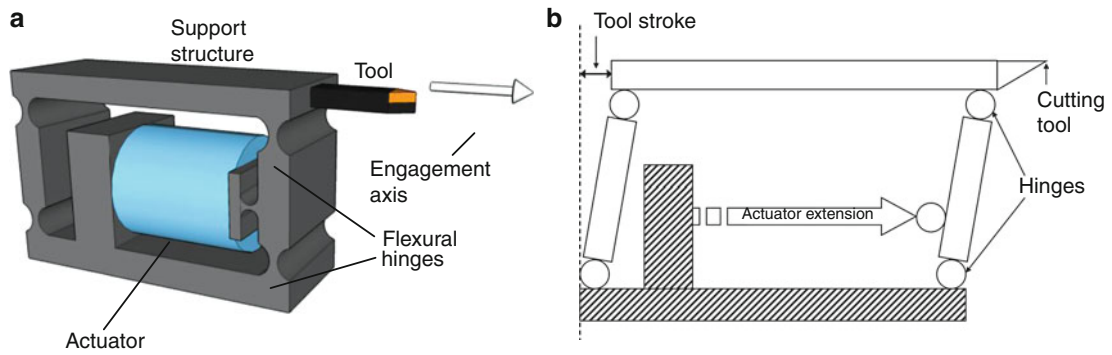
The stroke amplitude of the MPS used for VMT is generally in the range $50\text{--}100 \mu\text{m}$ with a bandwidth of $100\text{--}1,000 \text{ Hz}$.

The basic components of a MPS system are an actuator and a flexural support structure as shown in Fig. 3a, originally developed by (Hong and Ehmann 1995). This example MPS has a typical arrangement of the actuator in the flexural support structure. The extension of the actuator causes deflection of the structure at the hinges, which then moves the tool along the engagement axis towards the workpiece surface. Figure 3b shows a schematic representation of the tool stroke, which is exaggerated for visual clarity.

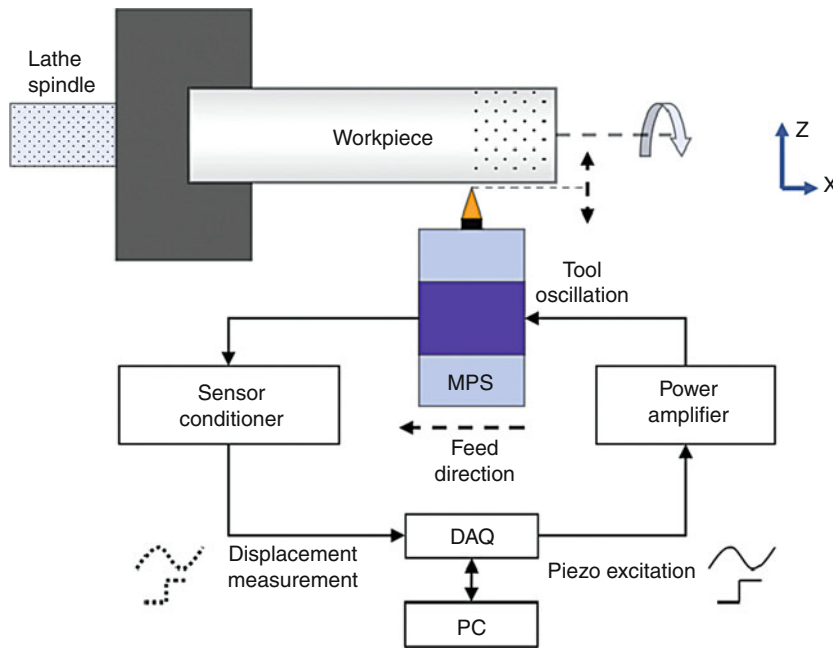
In order to achieve high-precision tool actuation, a piezoelectric actuator is typically used. These actuators can achieve nanometer-scale precision at a high response rate. The extension of a piezoelectric actuator is governed by the excitation voltage, which is computer controlled. For certain applications, a closed-loop control system is needed to achieve a high degree of precision.

Closed-loop Control

A closed-loop control system is commonly used for precision control of the MPS motion during VMT processing. As demonstrated in a previous section, each dimension of the texture layout depends on the amplitude of tool oscillation; therefore, controlling this parameter is essential to achieving accurate and consistent textures. Two main factors that cause inaccuracy in MPS actuation are, actuator hysteresis/drift and increased/inconsistent cutting forces. Drifting and hysteresis is common with piezoelectric driven actuators; however, for periodic motion the



Surface Texturing by Vibro Machining, Fig. 3 Micro-Positioning Stage (MPS) (a) typical MPS arrangement showing key components, (b) schematic of the MPS during actuator extension (not to scale)



Surface Texturing by Vibro Machining, Fig. 4 Schematic of the VMT closed-loop control system

effects are not as severe and are easily corrected by using a closed-loop control system. The force opposing the penetration of the tool into the workpiece surface is another source of inaccuracy. This is especially true for texturing of hardened metals, since the cutting force is proportional to the workpiece hardness. Increased cutting force impedes the penetration of the tool, preventing the tool from reaching the desired depth. For example, the associated error between predicted and actual texture dimensions for vibro-texturing heat-treated steel with a hardness of 55 HRC can be as much as 90% using an

open-loop control. However, for the same conditions a properly tuned closed-loop controller system is able to reduce this error to below 10%.

A typical MPS closed-loop control system consists of a computer interfaced data acquisition card with input and output analog terminals and a precision displacement sensor. The displacement sensor is mounted in such a way to measure the tool extension. The displacement measurement is read by the computer and compared with the set point. A typical proportional-integral-derivative (PID) control algorithm can be used to adjust the output

excitation voltage. Figure 4 shows a schematic of the closed-loop control system.

Key Applications

Surface Texture Processing in a Traditional Manufacturing Setting

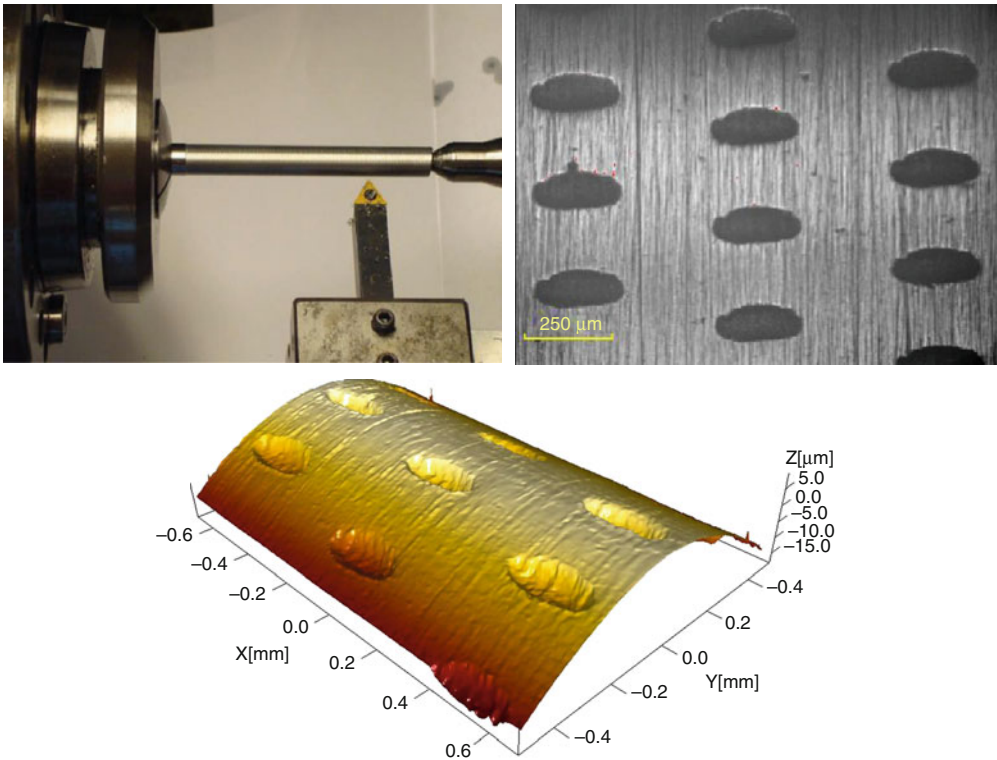
The VMT method utilizes a standard lathe and simple control system, which can be easily incorporated into a traditional manufacturing setting. The efficiency of the process makes it a cost effective alternative to other texturing techniques. The VMT method can be implemented

through a simple retrofit of a standard lathe by replacing the static tool mount with an MPS. Texturing can be performed inline with existing machining steps, eliminating the time consuming step of refixturing the workpiece in a separate machining setup.

A standard VMT procedure has two main steps. The first step is to establish the initial tool displace, Δ . This parameter is explained previously in reference to Fig. 2. To establish Δ , a single turning pass is taken over the entire texturing area with the MPS engaged at the Δ length. When the actuator is disengaged, the tool returns to the neutral position which is separated from the cutting tip by a length Δ . This first step also serves to eliminate any run-out that might exist between the tool and the workpiece surface. Considering the micro-scale size of the dimple feature, a small amount of run-out will cause inconsistent texturing. This process of establishing a uniform Δ also allows for the use of a standard lathe rather than an ultra-precision lathe. The second step in the VMT process involves a second turning pass, in which, the MPS is oscillated at desired wave characteristics. This set is responsible for the formation of the individual dimples.

Surface Texturing by Vibro Machining, Table 1 Dimple dimensions suitable for the VWT process

Depth	1–20 μm
Circumferential diameter	50–500 μm
Axial diameter	100–300 μm
Coverage density	5–25%



Surface Texturing by Vibro Machining, Fig. 5 Texturing a shaft surface with elliptical shaped dimples. Top left, the workpiece rod mounted in the lathe, top left, magnified view of the textures, and bottom, 3D reconstruction of the textured surface scanned using an optical interferometer. Note the depth is shown in microns

A finishing step typically follows the texturing process in order to remove any material bulges that may have formed around the dimples.

Textures on Cylindrical Surfaces for Journal Bearing Design

The VMT process can be used to produce rounded dimples or continuous grooves on a workpiece surface, especially on cylindrical surfaces. Both the outer surface of the shaft and inner surface of a bearing of a journal bearing set can be conveniently textured via the VMT process. As stated above, the dimensions of the texture layout are determined by the parameters of the machining process. The limits of the texture dimensions are related to the limits of the MPS, size of the workpiece, and shape of the cutting tool. Some common texture dimensions are shown in Table 1. These dimensions are consistent with dimple designs used for bearing applications.

The axial and circumferential diameters of the dimple are independent from one another; therefore, the shape of the dimple can either be circular or elliptical. As shown in Fig. 5, elliptical dimples are produced with the major diameter along the axial direction. For this case, if the frequency of the oscillation were reduced, the circumferential diameter would increase. A frequency could be calculated that would result in circular shaped dimples for the same setup.

Example. The shaft shown in Fig. 5 was textured with elliptical dimples with the major diameter oriented along the axial direction. The dimples have an axial diameter of 240 μm and a circumferential diameter of 100 μm . The frequency of oscillation was 120 Hz. If instead a circular dimple were desired (i.e., axial diameter=circumferential diameter) with the same depth, what frequency would be used, if all other machining parameters remained constant?

Solution. Considering (5) the circumferential diameter is inversely related to the frequency of oscillation. Since all other machining parameter are kept constant the equation can be reduced to $d_c=Y/f$, where Y is a constant representing the other machining parameters. Therefore, to calculate the frequency, f_{240} , needed to achieve 240 μm circumferential diameter dimples: $f_{240}=120*100/240=50$ Hz.

In addition to texturing cylindrical surfaces, the arrangement of the MPS rotation can be made so that the engagement axis is parallel to the lathe spindle axis. This setup is consistent with a standard facing process where the flat end face of the workpiece is machined. This arrangement

allows for the texturing of flat surfaces of components, such as those of thrust bearings and face seals.

References

- A. Greco, S. Raphaelson, K. Ehmann, Q.J. Wang, C. Lin, Surface Texturing of Tribological Interfaces Using the Vibromechanical Texturing Method. *J. Manuf. Sci. Eng.* **131**, 061005 (2009). doi:[10.1115/1.4000418](https://doi.org/10.1115/1.4000418)
- M.S. Hong, K.F. Ehmann, Generation of engineered surfaces by the surface-shaping system. *Int. J. Mach. Tool. Manufact.* **35**(9), 1269–1290 (1995)
- H.-S. Kim, E.-J. Kim, Feed-forward control of fast tool servo for real-time correction of spindle error in diamond turning of flat surfaces. *Int. J. Mach. Tool. Manufact.* **43**(12), 1177–1183 (2003)
- R. Montesanti, High Bandwidth Rotary Fast Tool Servos and a Hybrid Rotary/Linear Electromagnetic Actuator. Department of Mechanical Engineering, Boston, MIT, Ph.D. p. 555 (2005)
- Y.J. Noh, Y. Arai, M. Tano, W. Gao, Fabrication of Large-area Micro-lens Arrays with Fast Tool Control. *Int. J. Precis. Eng. Manuf.* **9**(4), 32–38 (2008)
- D.L. Trumper, X. Lu, Fast Tool Servos: Advances in Precision, Acceleration, and Bandwidth. F. Kimura, H. Kenichiro. (eds.), *Towards Synthesis of Micro-/Nano-systems*. The 11th International Conference on Precision Engineering, (Springer, London). pp. 11–19 (2007), ISBN: 978-1-84628-558-5

Surface Texturing for Non-conformal Surfaces

- [Surface Texture for Bodies in Non-Conformal Contacts](#)

Surface Topography

- [Topography of Engineering Surfaces](#)

Surface Topography Modification

- [Surface Texture for Bodies in Non-conformal Contacts](#)

Surface Treatment

- [Tribological Coatings for High-Temperature Applications](#)

Surface Variation in Tribological Processes

S. ILINCIC, G. VORLAUFER, F. FRANEK, A. PAUSCHITZ
Austrian Center of Competence for Tribology AC²T
research GmbH, Wiener Neustadt, Austria

Synonyms

Wear induced change of topography

Definition

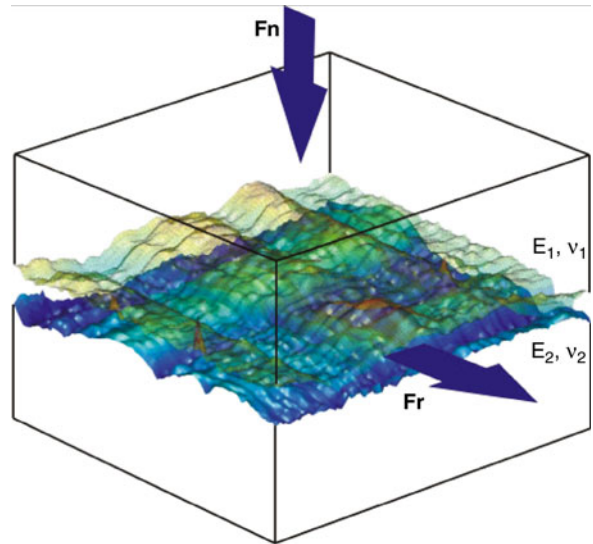
Tribological processes take place as two contacting surfaces are moving relative to each other, and both physical as well as chemical changes occur depending on the actual local conditions (considered as input data, such as surface topography, involved materials, and lubricants). Tribological processes are functions of time. They may change the surface geometry and the material composition, resulting in energy-related output effects: wear, friction, temperature, sound, and dynamic behavior. And, vice versa, the tribological processes may be changed by the resulting surface and contact conditions. The main effects of tribological processes are wear and friction, which are influenced by surface variation.

Scientific Fundamentals

Multi-asperity Contact

The surfaces of solids contain geometrical imperfections at many scales of length that influence tribological processes. At the atomic scale, surface irregularities are deviations from an atomically smooth or nominally flat surface. On the micro- and macro-scale, surface irregularities, depending on the length scale, are called roughness, waviness, or form error. These geometric imperfections are typically of random nature and influence, among others, the pressure distribution, the contact region, and subsurface stresses. Thus, in most cases variations of the surface roughness parameters, such as R_a , R_q and R_z , affect directly friction, wear, and lubrication processes (Bhushan 1999).

Due to roughness, contact between two surfaces occurs only at discrete spots that sustain the total compressive and tangential load, as depicted in Fig. 1. For a given load, the size of the contact spots depends on the mechanical properties and the surface topography of the contacting solids. The material properties include elastic modulus, yield pressure, and hardness. Some important

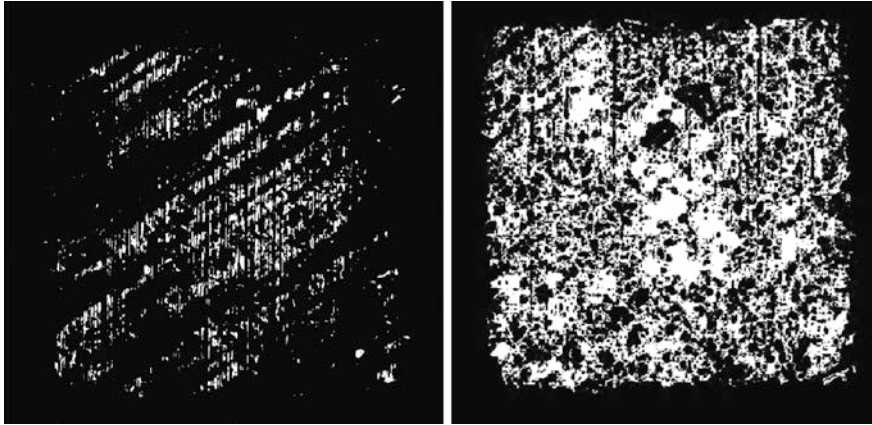


Surface Variation in Tribological Processes, Fig. 1 Two rough surfaces in contact

surface topography parameters are asperity distribution, tip radius, standard deviation of asperity height, and slope of asperity (Stolarski 2000).

With increasing normal load, contact spots also increase in size, and two or more asperities that were not in contact initially may coalesce and form a larger contact area (see Fig. 2). The pressure distribution between contacting solids can be quite irregular and complex. In addition, when the load is increased, there are no set rules that the contact spots follow. The “real contact area,” which is all created multi-asperity contact spots, is typically only a small fraction of the nominal contact area. In order to calculate the real contact area and pressure distribution (which serve as input for other analyses, e.g., wear estimation), it is necessary to employ either statistical methods as discussed in (Greenwood and Williamson 1966) or numerical techniques like described in Bhushan and Peng (2002).

Statistical techniques are commonly applied to predict some general trends in the case of rough surfaces, but they cannot provide detailed spatial distributions of contact stresses and deformations. These distributions are important to estimate local stresses and energy flow, and hence wear. Therefore, it is necessary to solve the multi-asperity contact problem numerically. Suitable numerical simulation techniques, e.g., the boundary element method, result in more accurate and detailed information about the contact stresses and the displacements in local contact regions.



Surface Variation in Tribological Processes, Fig. 2 Calculated contact area of a rough surface under two different loads (two-fold increase from left to right picture)

Wear Progress

Wear is one of the irreversible processes that is observed as a gradual change of the dimensions and shape of bodies in contact. Therefore, the surface topography changes with time and these changes have to be taken into account for a realistic description of tribological processes. The contact pressure distribution, shape variation, size, and position as well as the real contact area are unknown functions of time. A hypothesis of adhesive wear was given by J.F. Archard, where the total wear volume is proportional to the real contact area times the sliding distance (Archard 1957). A detailed list of the theoretical wear models is shown in Goryacheva (1998).

Archard introduced a coefficient K , which is the proportionality constant between real contact area A_r , sliding distance s , and wear volume V . Thus, Archard's law can be written as

$$V = K \cdot A_r \cdot s = K \cdot \frac{W}{H} \cdot s \quad (1)$$

where W is the applied load, H is the hardness of the softer surface, s is sliding distance, and K is a dimensionless proportionality constant, referred to in the literature as wear coefficient. The wear coefficient can also be imagined as the proportion of asperity contacts resulting in wear (Stachowiak and Batchelor 2001) and depend on the type of materials, interfacial friction, and the geometry of the surfaces. The value of K is obtained experimentally and usually has a small numerical value (e.g., for steel sliding against steel the order of magnitude is 10^{-8}). However, Archard's linear wear equation takes into account only global quantities. In order to study surface topography variation, a modified wear law that takes into account

locally varying friction energy density is used. The friction energy density as seen by each point of the contact interface at a given time t is given by

$$E(x, y, t) = \int_0^t \mu(x, y, \tau) \cdot p(x, y, \tau) \cdot v(x, y, \tau) \cdot d\tau \quad (2)$$

where μ is the coefficient of friction, p the contact pressure, and v the sliding velocity. Typically, the coefficient of friction is known only globally and determined experimentally. Knowing the local energy density, the wear depth at each point (x, y) of each contact body B ($B \in \{1, 2\}$), and the time t is then given by

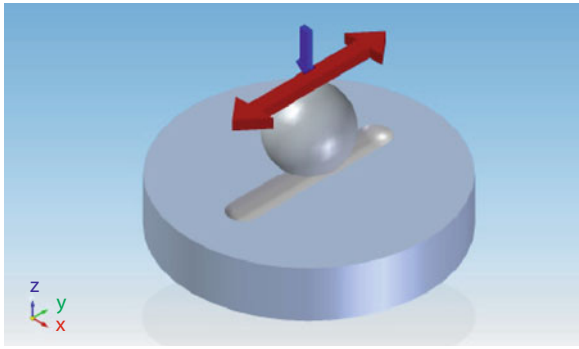
$$h^B(x, y, t) = \frac{K^B \cdot E(x, y, t)}{H^B} \quad (3)$$

where K^B is the wear coefficient of contact body B that is obtained through experiments and H^B is the hardness of each body. Using the modified wear law (3), it is possible to determine the local wear depth at any given time. The local wear modifies the surface topography, which in turn has an influence on the real contact area, contact pressure distribution, and friction energy density.

Key Applications

Experimental Ball-On-Plate Reciprocating Sliding Wear Test

The reciprocating sliding wear test is suitable as a model test for a wide range of tribological applications. The contact may be either dry or lubricated. Since the contact situation is simple and consists of a ball pressed against a plate (see Fig. 3), it is frequently used to



Surface Variation in Tribological Processes, Fig. 3 Ball-on-plate wear test setup (schematic)

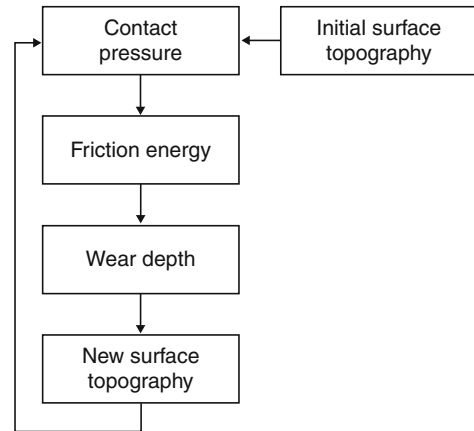
investigate friction and wear behavior of materials, surface coatings and lubricants. For some routine applications, standardized test rigs and testing procedures exist (e.g., linear-oscillation (SRV) test machine, high frequency reciprocating rig (HFRR)).

The progressive loss of substance that occurs on test specimens is often monitored online by measuring the relative approach of the two counteracting bodies. This relative approach may be regarded as an average wear height. However, one has to keep in mind that the relative approach may be influenced by thermal expansion of the solids and build up of wear debris in the contact zone. Furthermore, since the real contact area is changing with progress of wear, measurement of the current relative approach in general cannot be used as an accurate measure of the worn volume. After completion of the test, the total worn volume of the test specimens can be accurately determined, e.g., using optical profilometry (see ► [Wear Quantification by Comparison of Surface Topography Data](#)). Applying (1) or (3), the modified Archard wear coefficients can be determined.

Computer Simulation of Wear Progress in a Ball-On-Plate Oscillating Sliding Wear Test

Using (1), (2), and (3), the time dependence of wear may be numerically simulated. As shown in Fig. 4, each iteration (i.e., each time step of the simulation) consists of the following steps:

1. Calculation of real contact pressure distribution and real contact area taking into account the surface topography of the counteracting bodies;
2. Calculation of the frictional energy density (2) for the current time step;
3. Calculation of the local wear height (3);
4. Update of the surface topography.



Surface Variation in Tribological Processes, Fig. 4

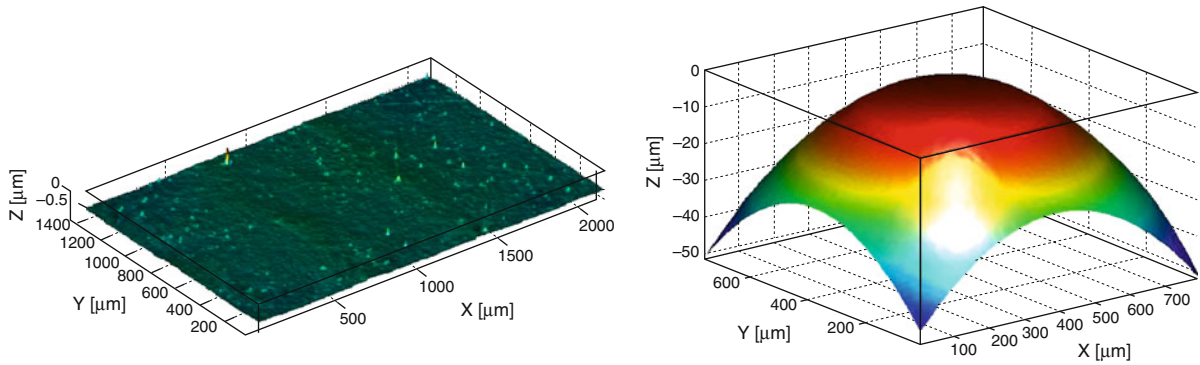
Schematic flow chart diagram of wear simulation (not considering additional micro-effects due to local thermal expansion and presence of wear debris)

Steps 1–4 are repeated sequentially in a loop, thereby simulating wear over a given time interval.

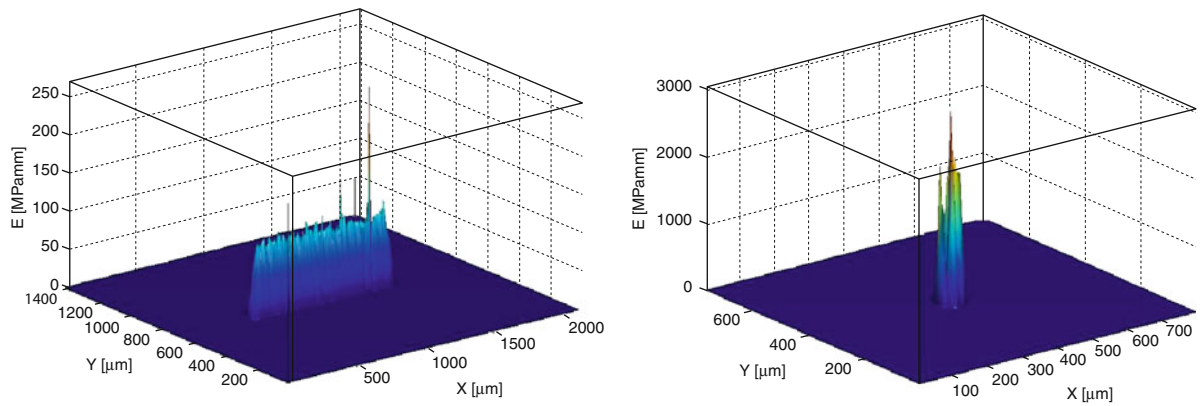
From the computational point of view, step 1 is challenging both in terms of the applied numerical technique and in terms of computational costs. This step essentially consists of solving the multi-asperity contact problem. A combined finite element–boundary element approach that has been successfully applied to the multi-asperity contact problem is described in Ilincic et al. (2009).

Example. The simulation has been run using a real world example consisting of a ball with a diameter of 6 mm made of 100Cr6 (AISI E-52100) steel with a hardness 58–66 HRC and a plate of the same material with much lower hardness of 190–200 HV (Ilincic et al. 2009). Their initial topographies (see Fig. 5) were measured by a commercial grade confocal microscope. Normal load was 2 N and oscillating motion was 1 mm (peak to peak) with a frequency of 50 Hz. The wear coefficients were obtained from experiments as described above and are $K^B = 1.434 \times 10^{-9}$ for the ball and $K^P = 1.001 \times 10^{-9}$ for the plate. Test duration was approximately 80 min. The coefficient of friction was experimentally determined as $\mu = 0.32$. Figure 6 shows the energy density distribution on plate and ball after the first sliding cycle ($t = 0.02$ s).

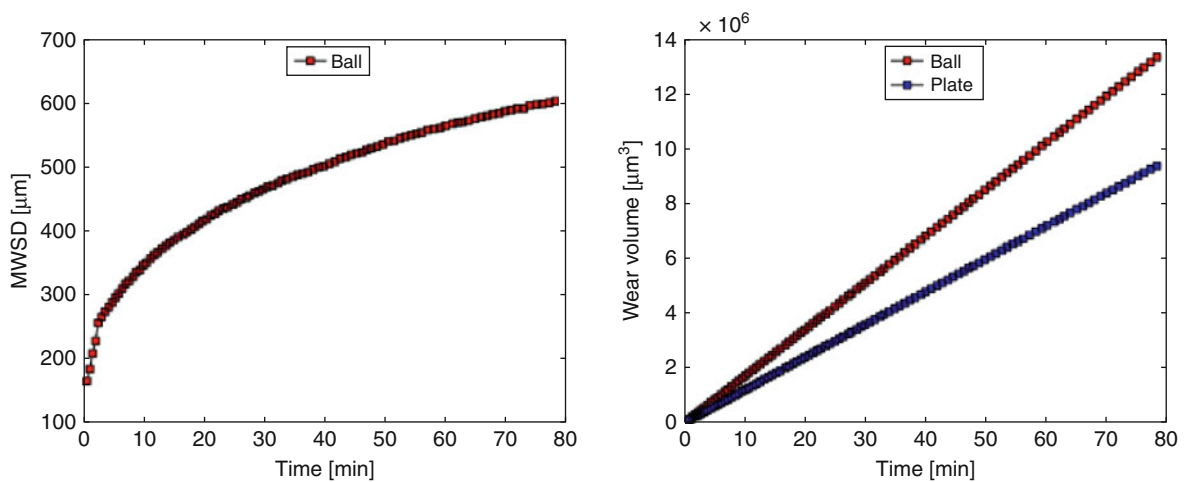
During the simulation the lateral dimensions of the regions affected by wear (i.e., the wear scar) increase as well as the wear depth. Since the wear scar has an almost circular shape on the ball it is usually represented by an average diameter. The wear volume is calculated by



Surface Variation in Tribological Processes, Fig. 5 Measured surface topography of plate (left) and ball (right)



Surface Variation in Tribological Processes, Fig. 6 Friction energy density on plate (left) and on ball (right) after 0.02 s of sliding



Surface Variation in Tribological Processes, Fig. 7 Mean wear scar diameter on ball (left) and wear volume of ball and plate (right), respectively, as a function of test time

integration of the wear height over the wear scar. The results of the simulation can be summarized as shown in Fig. 7 by plotting the wear scar diameter of the ball and the wear volume of plate and ball over time. It must be stated that, by assuming constant wear rates, the effect of running in is ignored and a linear relation of wear volume over time is achieved as expected from (1). The numerical simulation results show an overall very good agreement with experimental test data (Hunger et al. 2008).

Cross-References

- [Confocal Microscopy](#)
- [Contact of Rough surfaces: The Greenwood and Williamson/Tripp, Fuller and Tabor Theories](#)
- [Fractal Contact Mechanics](#)
- [Sliding Wear](#)
- [Stochastic Contact Theories: Theories of Surface Roughness and Applications to Contact Mechanics](#)
- [Wear Quantification by Comparison of Surface Topography Data](#)

References

- J.F. Archard, Elastic Deformation and the Laws of Friction. *Proc. R. Soc. A* **327**, 190–205 (1957)
- B. Bhushan, *Handbook of Micro/Nano Tribology* (CRC Press, Boca Raton, FL, 1999). Chap. 4
- B. Bhushan, W. Peng, Contact mechanics of multilayered rough surfaces. *ASME Appl. Mech. Rev.* **55**(2002), 435–480 (2002)
- I.G. Goryacheva, *Contact Mechanics in Tribology* (Kluwer Academic, Dordrecht, 1998), p. 197
- J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces. *Proc. R. Soc. A* **295**, 300–319 (1966)
- H. Hunger, U. Litzow, S. Genze, N. Dörr, D. Karner, C. Eisenmenger-Sittner, Tribological characterisation and surface analysis of diesel lubricated sliding, in *16th International Colloquium Tribology – TAE, Stuttgart* (2008), p. 85
- S. Ilincic, G. Vorlaufer, P.A. Fotiu, A. Vernes, F. Franek, Combined finite element – boundary element method modelling of elastic multi-asperity contacts. *Proc. IMechE Pt. J: J. Eng. Tribol.* **223**(5), 767–776 (2009)
- G.W. Stachowiak, A.W. Batchelor, *Engineering Tribology* (Butterworth-Heinemann, Boston, 2001), p. 447
- T.A. Stolarski, *Tribology in machine design* (Butterworth-Heinemann, Oxford, 2000), p. 14

Surface Waviness

- [Contact with Micro Vibrations for Friction Control and Wear Reduction](#)

Surface-Originated Pitting

- [Probabilistic Life Prediction Models for Rolling Contact Fatigue](#)

Surmount Clutch

- [Overrunning Clutch](#)

Suspension Assembly for Disk Drive

- [Suspension Assembly for Hard Disk Drive](#)

Suspension Assembly for Hard Disk Drive

MOHAMMAD R. KAZEMI

Tyco Thermal Controls, Menlo Park, CA, USA

Synonyms

[Suspension assembly for disk drive](#); [Suspension for disk drive](#); [Suspension for hard disk drive](#)

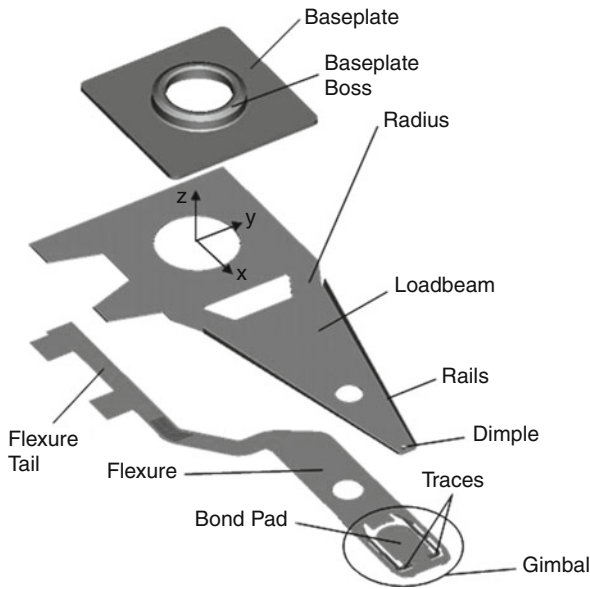
Definition

A suspension assembly is a component of the hard disk drive (HDD) that positions the slider over the surface of a rapidly spinning disk.

Scientific Fundamentals

Introduction

Disk drives store and retrieve information by using magnetic heads to write onto and read from rapidly spinning disks. A magnetic head is attached to a slider, and the slider is held by a suspension assembly. The assembly of a suspension and a slider forms a head gimbal assembly (HGA), and the assembly of an HGA (one or several) and



Suspension Assembly for Hard Disk Drive, Fig. 1 Exploded view of a typical suspension assembly

an actuator arm constitutes a head stack assembly (HSA). [Figure 1](#) shows the exploded view of a suspension assembly. As can be seen, a suspension assembly consists of three main components that are welded to each other:

- Base plate
- Load beam
- Flexure

A suspension assembly is attached to the actuator arm, typically through the swaging process at the baseplate boss (see (► [ABS Designs](#)) for more information on suspension terms). The loadbeam is rolled down at radius region so that the suspension assembly can provide a gram load force (in the negative z-direction) when an HSA is integrated with a set of disks. The gram load force (F_z) is transferred to the slider through a formed, semi-spherical feature, named a dimple. That is why the location of the dimple on a suspension assembly is referred to as the “load point.” Another formed feature on the loadbeam is the rail, which provides increased rigidity for the loadbeam.

Flexure is made of a laminate material that is composed of stainless steel (SST), dielectric, and copper. The front portion of the flexure is called the gimbal and the back portion is referred to as the flexure tail. The gimbal is the major mechanical portion of the flexure that controls the pitch and roll stiffness as well as pitch and roll static attitudes. The flexure also includes

electrical traces (the copper layer of the laminate flexure) that transmit signals between the magnetic head and pre-amp. The slider is attached to the bond pad portion of the gimbal. The bond pad is in contact with the dimple and can freely pivot around the dimple for the pitch and roll rotations (y-direction and x-direction rotations, respectively).

A typical suspension assembly has six traces: two for reading, two for writing, and two for thermal FH control (TFC). The latter pair is used to supply electrical current to a dedicated electrical resistance that is deployed for heating the slider in the proximity of magnetic head. Such heating results in thermal expansion of the heated area and, consequently, the protrusion of the magnetic head towards the disk surface. This phenomenon allows for lowering the FH of a magnetic head with no need to fly the slider lower. In addition, the TFC allows HDD original equipment manufacturers to compensate for FH variations due to (a) manufacturing tolerances, (b) ambient condition changes, and, (c) write-induced pole tip protrusion (PTP).

The suspension assembly enables the magnetic head to fly just above the surface of the spinning disk with a flying height (FH) of only a few nanometers. In addition, the suspension assembly also needs to precisely position the magnetic head on the track so that data can be accessed from that track. However, the rotation of the disk induces airflow, which, in turn, causes the off-track vibration of the suspension and, consequently, track misregistration (TMR) of the magnetic head. To this end, more robust suspension design is needed to lower the magnitude of such vibration.

Suspension Loads

When an HSA is integrated with a set of disks, the following loads are applied on the slider by the suspension assembly:

1. Gram load force (F_z)
2. Pitch torque (M_y)
3. Roll torque (M_x)

The gram load is due to fact that the rolled-down suspension is pushed back up when the HSA is loaded into disk assembly. The rate of change in the gram load force versus the z-direction displacement of the dimple is referred to as spring rate (SR). The spring rate, and consequently gram load force, is mainly controlled by the geometry of the radius region. For a typical suspension assembly, the spring rate is between 10 and 25 N/m and the gram load force is between 0.5 and 3.0 gf.

Pitch torque is a mechanical torque in the y-direction that is mainly driven by the gimbal portion of the suspension assembly. Pitch torque can be expressed as the product of pitch stiffness (K_P) and pitch static attitude (PSA):

$$M_y = K_P \times \text{PSA} \quad (1)$$

The pitch stiffness is the rotational stiffness of the gimbal in the y-direction and presently ranges between 0.5 and 1.2 $\mu\text{N.m/deg}$. PSA refers to the angle of the gimbal bond pad relative to the disk surface in the y-direction and presently ranges between 0° and 2.5° (see (► [Head Disk Interface for Patterned Media](#)) for sign convention for PSA). However, the suspension assemblies that are designed for load/unload (L/UL) disk drives typically have a PSA larger than 0.5° .

Roll torque is a mechanical torque in the x-direction that is mainly controlled by the gimbal portion of the suspension assembly. Similar to pitch torque, roll torque is also expressed as the product of roll stiffness (K_R) and roll static attitude (RSA):

$$M_x = K_R \times \text{RSA} \quad (2)$$

In this equation, K_R and RSA are the rotational stiffness of the gimbal in the x-direction and the angle of the gimbal bond pad relative to the disk surface in the x-direction, respectively (see (► [Head Disk Interface for Patterned Media](#)) for sign convention for RSA). Similar to pitch stiffness, roll stiffness ranges between 0.5 and 1.2 $\mu\text{N.m/deg}$. However, RSA is typically zero for most of the present suspension assemblies.

Key Applications

Flying Performance

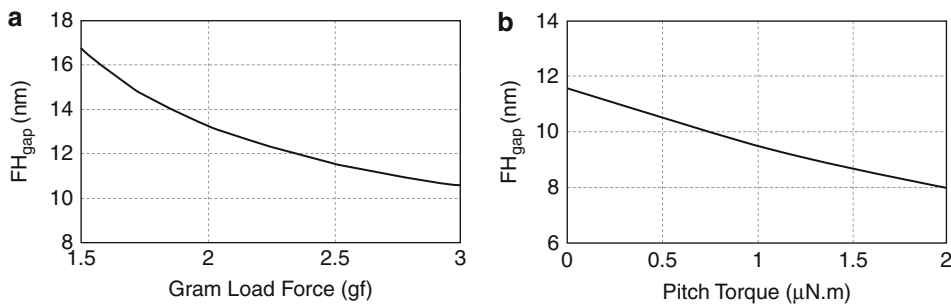
The gram load force as well as the pitch and roll torques are the loads that a suspension assembly imposes on a slider. During a track following process, those loads are

balanced by the counter loads supported by the airflow between the airbearing surface (ABS) of the slider and the disk surface. Hence, for a specific set of suspension loads (F_z , M_x , M_y), the slider can only acquire a specific set of flying characteristics that allow for the balance of the suspension loads with the airbearing loads. Those flying characteristics are flying height of the magnetic head (FH_{gap}), dynamic pitch angle (DPA), and dynamic roll angle (DRA).

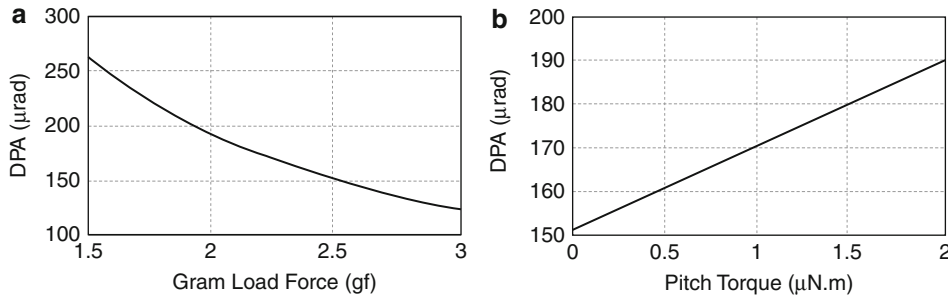
Among those flying characteristics, FH_{gap} and DPA are strongly affected by the gram load force and pitch torque, while DRA only shows strong correlation with the roll torque. The variation of flying characteristics with the suspension loads are shown in Figs. 2–4. As can be seen, FH_{gap} drops with the increase of gram load force and pitch torque. In addition, the DPA also declines when the gram load force rises, but it climbs when the pitch torque increases. Similarly, DRA increases with the increase of the roll torque.

Disk drive manufacturers specify F_z , PSA, RSA, K_P , and K_R for a suspension design in an HDD program. Suspension manufacturers then design the suspension in compliance with those specifications as well as other specifications (resonance performance, windage performance, shock performance, hygrothermal performance, etc.). Meeting those suspension specifications allows HDD manufacturers to achieve the target flying characteristics. However, manufacturing tolerances at suspension, HGA, and HSA levels cause variation of those suspension specifications.

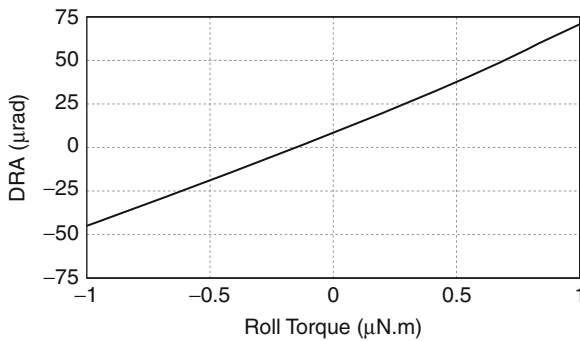
For many years, manufacturing precisions at various levels were being constantly improved, allowing for tighter tolerance of the suspension specifications and, consequently, tighter variation of flying characteristics. However, the advent of TFC in the HDD industry has dramatically softened the demand for those tolerance-tightening efforts. That is because the variation of FH_{gap}



Suspension Assembly for Hard Disk Drive, Fig. 2 Plots of flying height at magnetic head (FH_{gap}) versus gram load force (a) and pitch torque (b). The data correspond to a typical Pemto slider



Suspension Assembly for Hard Disk Drive, Fig. 3 Plots of dynamic pitch angle (DPA) versus gram load force (a) and pitch torque (b). The data correspond to a typical Pemto slider



Suspension Assembly for Hard Disk Drive, Fig. 4 Plot of dynamic roll angle (DRA) versus roll torque. The data correspond to a typical Pemto slider

can now be compensated for by adjusting the thermal protrusion of the head.

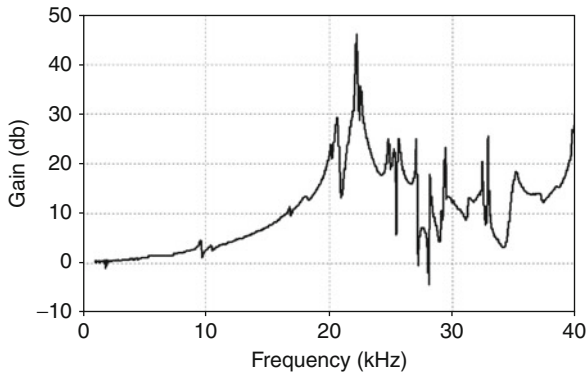
Mechanical Robustness

The natural frequencies and stiffness are the main factors that define the mechanical robustness of a suspension assembly. A suspension assembly that has higher natural frequencies supposedly indicates superior performance in the frequency response function (FRF) and windage tests (see [FRF & Windage Standard Definitions Document](#) for more information on FRF and windage tests)). Among the natural frequencies of a suspension assembly, sway frequency (f_s) and first and second torsional and bending frequencies (f_{T1} , f_{T2} , f_{B1} , f_{B2}) are the ones that contribute the most into the mechanical robustness of a suspension

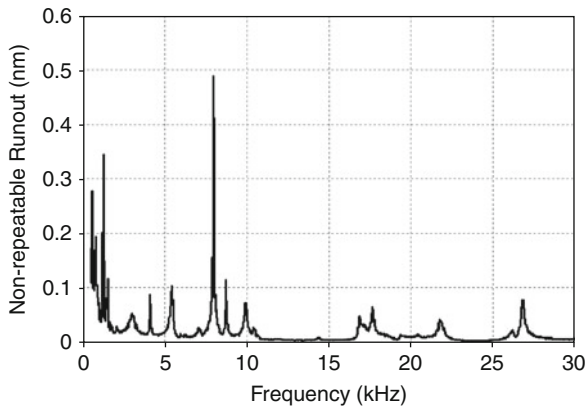
assembly. The processing of the radius region also has a noticeable impact on suspension performance. In particular, the rolling location in the radius region significantly impacts the gain of the first torsion mode in the FRF test. To this end, the rolling location is tuned, targeting minimal gain of the first torsion.

The lateral stiffness (linear stiffness in the y-direction), the torsional stiffness (rotational stiffness in the x-direction), and the spring rate (linear stiffness in the z-direction) of a suspension assembly also play a significant role in the mechanical robustness of a suspension assembly. In general, higher stiffness is desired because it results in lower amplitude of vibrations of a suspension assembly due to the excitations induced by the voice coil motor (VCM) or airflow. However, higher spring rate causes larger variations of the gram load force due to z-direction tolerance of the HSA. Larger gram load force variation results in larger tolerances of natural frequencies of the suspension assembly which is undesired.

Figure 5 shows the results of the FRF test for a typical suspension assembly. In this test, the suspension assembly is shaken in the y-direction at the baseplate and the slider vibration in the y-direction is recoded as the response function. As can be seen, the sway mode (at 22 kHz) strongly manifests itself in this plot. The result of the windage test for a typical suspension assembly is shown in Fig. 6. This test represents the off-track (y-direction) vibration of the slider induced by the airflow that, in turn, is caused by the disk rotation. It can be observed that the first torsion mode (at 8 kHz) exhibits the largest spike in this plot. The first bending mode (at 5.5 kHz) also appears in this plot due to the fact that the suspension assembly is slightly twisted because of the manufacturing imperfections.



Suspension Assembly for Hard Disk Drive, Fig. 5 Plot of gain versus frequency obtained from the FRF test of a typical suspension assembly



Suspension Assembly for Hard Disk Drive, Fig. 6 Plot of non-repeatable runout versus frequency obtained from the windage test of a typical suspension assembly

The servo system helps the magnetic head to remain on track by using the positioning data, provided by the servo bursts, to actuate the HSA by VCM. However, in some disk drives, a secondary actuation system is used to achieve better track-following performance. The secondary actuation system is typically driven by a pair of piezoelectric motors. Those motors can be mounted on arm, baseplate, loadbeam, gimbal or slider. A suspension assembly with the secondary actuation system is referred to as a dual stage actuator (DSA) suspension (see ([Glossary of Suspension Terms Purpose](#)) and ([Dual Stage Actuator](#)) for more details on dual stage actuation).

In some suspension assemblies, a damper is added as a separate component so that the gains of dominant modes and the vibration amplitudes can be suppressed.

In a typical configuration, the damper consists of a viscoelastic layer and a stainless steel layer. The damper is commonly attached to the center of the loadbeam and may extend into the radius region as well. Similarly, some HDD manufacturers tend to accommodate a damper on their actuator arms so that the gains of the arm modes are damped. The deployment of a damper in a suspension assembly comes with some side effects. In particular, a damper degrades the shock performance and increases the sensitivity of suspension characteristics to the ambient temperature and humidity.

Future Challenges

Disk drive manufacturers are planning to demonstrate an areal density of 10 Tb/in.² in 2016. Such areal density poses significant challenges for the design of magnetic heads, media, signal processing, servo, and tribology. Various major technologies are currently under investigation for achieving such areal density. Heat-assisted magnetic recording (HAMR), bit-patterned magnetic recording (BPMR) ([Sign Convention Standard](#)), and two-dimensional magnetic recording (TDMR) are examples of those technologies. For the tribology, the requirement for 10 Tb/in.² comes down to lowering the FH_{gap} to about 1 nm. Given that the glide avalanche is not expected to fall below 0.75 nm, the clearance between the magnetic head and disk surface should be lowered to 0.25 nm with a maximum allowable modulation of about ± 0.025 nm. Such low values for the nominal and modulation of clearance pose a significant challenge to the slider flying performance.

Maintaining 0.25 nm clearance during the track-following process requires a more enhanced FH control technology (TFC or other methods). In particular, an adaptation of the FH control that allows for the dynamic compensation of small FH_{gap} variations may be necessary. Such an enhanced FH control system is likely to require additional flexure traces so that the FH_{gap} can be continuously sensed. The sensed FH_{gap} data can then be used for a closed-loop FH control system.

Since intermittent head-disk contact is likely to occur at such low clearance, the magnetic head will experience off-track and down-track vibrations as well as FH_{gap} modulation ([► ABS Designs](#)). To this end, the mechanical robustness of the suspension needs to be significantly improved to minimize those oscillations. In addition, deployment of a secondary actuator is likely to be necessary for achieving better track-following performance. Furthermore, the deployment of HAMR or BPM may also pose some design change requirements for the suspension assembly. In particular, both HAMR and BPM

may require additional flexure traces while the HAMR may also cause a thermal problem for the flexure. Therefore, future suspension designs need to address those issues as well.

Cross-References

- [ABS Designs](#)
- [Head Disk Interface for Patterned Media](#)

References

- Dual Stage Actuator. The International Disk Drive Equipment and Materials Association (IDEMA), Doc. # H18-04
- FRF & Windage Standard Definitions Document. The International Disk Drive Equipment and Materials Association (IDEMA), Doc. # H17-04
- Glossary of Suspension Terms Purpose. The International Disk Drive Equipment and Materials Association (IDEMA), Doc. # H8-94
- Sign Convention Standard. The International Disk Drive Equipment and Materials Association (IDEMA), Doc. # H7-94

Suspension for Disk Drive

- [Suspension Assembly for Hard Disk Drive](#)

Suspension for Hard Disk Drive

- [Suspension Assembly for Hard Disk Drive](#)

Sustainable Technology for Tribological Textiles

THIRUMAL YASODHA

Department of pharmacology, Al-Tahadi University, Libya

Synonyms

[Bio-reinforced composites](#); [Ecofriendly textiles](#); [Green textiles](#); [Natural fiber composites](#); [Plant based natural fiber composites](#); [Polymer composites](#)

Definition

Strategies for sustainable technology through blending of natural fibers from natural resources with polymers and plastics. The uses of fibers in industry as unique tribologic

textiles include clothing, non-wovens, and technical textiles.

Scientific Fundamentals

Renewable resources that were important before the Industrial Revolution are regaining interest in our modern society because of their positive effects on various technologies, industries, the environment, and the economy. A significant advantage of renewable resources exists in their contribution to the conservation of finite fossil resources and their importance in mitigating the green house effect.

The universe of “bio-fibers” is fairly broad. Included are very short wood fibers from both deciduous and coniferous sources, which are used as fillers in extruded plastic lumber and molding compounds. From a commercial standpoint, the most viable structural fibers come from purpose-grown textile plants and some fruit trees.

Such fibers can generally be classified into three types. “Bast” fibers, such as flax, hemp, jute, and kenaf, are noted for being fairly stiff when used as a composite reinforcement. Leaf fibers, including sisal, henequen, pineapple, and banana, are noted for improving composite toughness with somewhat lower structural contribution. Finally, seed or fruit fibers – cotton, kapok, and coir (from coconut husks) – demonstrate elastomeric type toughness but are not structural (Drzal et al. 2005). Promoted as low-cost and low-weight alternatives to fiberglass, these agricultural products, including flax, jute, hemp, and kenaf, signaled the start of a “green” industry with enormous potential.

Among the purpose-grown plants, bast fibers represent the vast majority of natural fibers with potential for use in composites. Bast plant stems are characterized by long fibers surrounding a core of pulp or short fibers and covered with a protective bark layer. Separation of the useful fibers from the bark and core starts with a process called “retting,” in which the cut stalks are soaked in water or left in the field in a humid environment for up to several weeks to degrade the natural binders. This makes the fiber bundles easier to process by mechanical means, or by hand, as is the case in many developing countries. As of the mid-1990s, flax and jute were the principal fibers used in biocomposites, and they have been joined by higher-strength industrial hemp and kenaf, at least in automotive applications. There is wide potential for use of natural fibers in the automotive industry.

The costs added to the loss of good fibers during processing to fix the material price. It is important to emphasize that the good fiber yield (after processing) per

hectare is a more important figure than the tons biomass per hectare. This implies that the raw material must be adapted to specific processing, e.g., decortication and refining. The increase in the demand for plant materials is due to reduction of costs associated with technical processes.

Another parameter that is very important, especially in technical applications, is the coefficient of variation (CV). Unlike man-made fibers, the CV values for natural fibers are higher and therefore require more attention in processing. In general, narrowing the CV by selection means additional costs because of severe losses.

The use of naturally processed (bioprocessed) fibers in industry as unique tribologic textiles may be divided into three fundamental categories:

1. Textile applications (clothing)
2. Non-wovens
3. Technical textiles

For tribologic textile fibers to be cost competitive in the future, tailor-made, high-quality fibers must be provided for textile and non-textile markets. The following strategies are suggested:

1. Non-woody plant resources should be identified as alternatives to the present synthetic and natural fibers.
2. Suitable methodologies involving mechanical (harvesting), chemical, and biological (enzymatic and microbial) processes for retting and decortications should be determined.
3. Quality assessment of fibers using infrared spectroscopy, ultraviolet/visible spectroscopy provides information about supramolecular and morphological structures of pectin, lignin, and cellulose. Based on these details, retting degree and fineness can be determined.

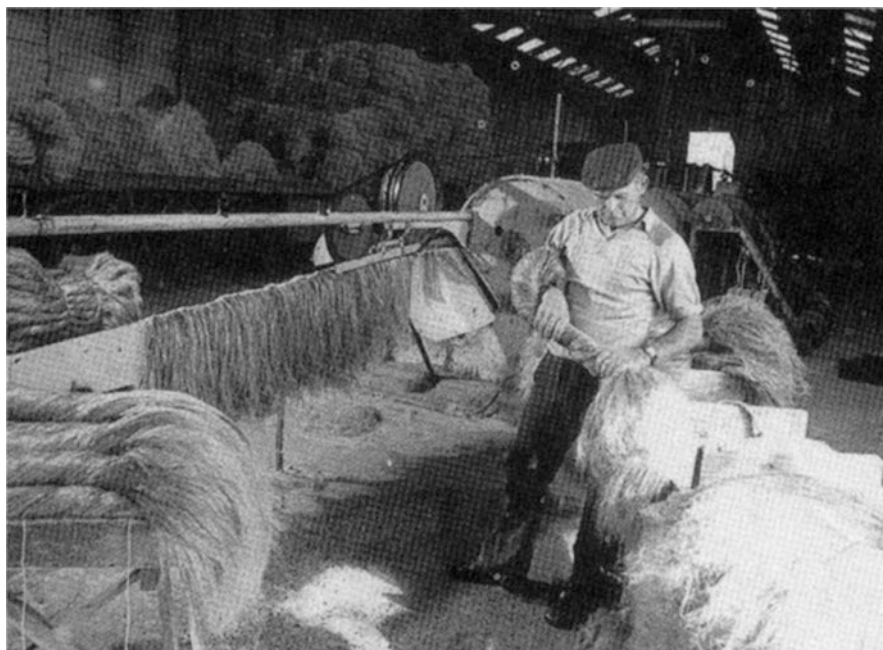
More than 20 families of fiber crops have been identified for textiles and composites, including Linaceae, Moraceae, Tiliaceae, Malvaceae, Urticaceae, Leguminosae, Sterculiaceae, Agavaceae, Amaryllidaceae, Liliaceae, Bromeliaceae, Musaceae, Palmae, Bombaceae, and Asclepiadaceae. All these plants consist of fibers that are long and parallel or in bundles running continuously from root through trunk, leaf stalk, or leaves. Agro-based fibers are classified as follows: (1) bast or stem fibers, which are the fibrous bundles in the inner bark of the plant stem running the length of the stem; (2) leaf fibers, which run the length of leaves; (3) seed-hair fibers; and (4) core, pith, or stick fibers, which form the low-density, spongy inner part of the stem of certain plants. Examples of bast or stem fibers include jute, flax, hemp, kenaf, ramie, roselle, and



Sustainable Technology for Tribological Textiles, Fig. 1
Flax harvesting (Collier and Tortora 2001)

urena. Leaf fibers include banana, sisal, henequen, abaca, pineapple, cantala, caroa, mauritius, and phormium. Seed-hair fibers include coir, cotton, kapok, and milk weed floss. Core fibers represent the center or pith fibers of such plants as kenaf and jute and can represent more than 85% of the dry weight of these plants. The remaining fibers include roots, leaf segments, flower heads, seed hulls, and short stem fibers. While individual single fibers in all of these classes are quite short (except for flax, hemp, ramie, cotton, and kapok), fiber bundles can be quite long (Figs. 1–3). For example, hemp, jute, and kenaf can have fiber bundles as long as 400 cm, and abaca, mauritius, and phormium are about half this length (Liese 1985; Prasad et al. 1995; Webber 1993). The fibers of the trunk and leaves of the date palm, *Phoenix dactylifera*, of Palmae are used as material for rope and basket fibers (Smith 1982).

Several scientific approaches have been made to isolate these fibers without breaking or lessening their strength. This involves processes such as retting, fiber separation, and fraction purification, which increases the demand for plant material and reduces the costs associated with fraction isolation. This also gives farmers a different option in their crop utilization (Rowell et al. 2000; Saheb and Jog 1999).



Sustainable Technology for Tribological Textiles, Fig. 2 Flax processing (Collier and Tortora 2001)



Sustainable Technology for Tribological Textiles, Fig. 3
Flax fiber

Fiber Grade Used in Tribologic Textiles

Fibers primarily used in hand weaving are classified into two groups: the linuan, or fine fibers, and coarse fibers. The Red Spanish or native variety is mainly used for hand weaving to produce valuable items such as handkerchiefs, table napkins, table cloths, fans, and gowns. To determine the fiber grade, various parameters such as tensile strength, fineness, moisture content, total cellulose, lignin, and residual gum are considered.

Blending Processes of Natural Fibers and Composites

The techniques used to manufacture biocomposites are based largely on existing techniques for processing plastics or composite materials such as spray-up, hand lay-up, resin transfer molding, closed-cavity bag molding, vacuum infusion molding, filament winding, injection molding, and compression molding. It is probably fair to say that the majority of current biocomposite materials based on thermoplastic polymers, such as polypropylene and polyethylene, are processed by compounding and extrusion.

1. Compounding and extrusion of thermoplastic polymers and natural fibers

During compounding, the thermoplastic polymer is heated, either by an external heat source or as a result of mechanical shearing in the extruder, so that it melts. In this state, wood fiber, usually in the form of flour, can be added along with other additives to improve the characteristics of the resultant material. Once the constituents have been thoroughly mixed, the compound can be either extruded directly in the final product or pelletized and packed as a precursor to further extrusion or injection molding processes. One of the current limitations of compounding and extrusion is that only relatively short fibers (which impart

limited reinforcement) can be used. If longer fibers are to be included, alternative methods may need to be employed (Hearle 2001).

2. Co-mingling of thermoplastic and natural fibers

In the automotive industry, longer fibers from flax, hemp, kenaf, and cotton are frequently used. These are generally mingled together with fibers of thermoplastic polymer used as a reinforcement to form a non-woven “fleece” that is hot pressed to melt the thermoplastic fiber, thereby forming the composite. The advantage of this approach is that longer fibers (with better reinforcement potential) can be used.

3. Processing of thermosetting polymer matrix composites

Although there is significantly less commercial production of thermosetting polymer matrix biocomposites, interest in this area remains high. Manufacturing techniques broadly mirror those found in the “traditional” composite industry.

I. Open Mold Processes. These processes include spray-up hand lay-up molding in one-sided molds. These are low-cost processes and are commonly used for making boat hulls and decks, RV components, truck cabs and fenders, spas, tubs, showers, and other fiberglass composite products.

In a spray-up application, the mold is waxed, sprayed with gel coat, and then cured in a heated oven at 120°F. After the gel coat cures, the mold is sprayed with a mixture of catalyzed resin. Fiber is chopped in a hand-held gun and fed into a spray of catalyzed resin. The spray-up is rolled out so the laminate can be compacted. Wood, foam, or other core material may then be added. A secondary spray-up layer imbeds the core between the laminates (sandwich construction). The part is then cured, cooled, and removed from the reusable mold.

Since the introduction of nylon, the first manufactured synthetic fiber, in the 1940s, synthetic fibers have had a significant impact on the quality of our lives and many consumer products. Demand for natural fibers continues to increase, however, because of their many outstanding properties, including aesthetics, comfort, and biodegradability. Farmers, fiber producers, and scientists worldwide have been exploring the use of alternative fiber crops (kenaf, jute, and hemp), crop residues, and agricultural by-products, which are often underutilized and undervalued. For example, kenaf fibers are biodegradable, environmentally friendly, and inexpensive to grow; plus, they will grow almost anywhere and in any type of soil (Romanoschi et al. 1997; Cooke 1990).

Jute also is a relatively cheap, easy-to-grow fiber having good mechanical properties. Flax and ramie, textile fibers

used for long periods of time in different parts of the world, became cost competitive because of new developments in the fiber extraction process.

II. Closed Mold Processes. This technology is used to produce precision parts in a variety of industries, especially for applications requiring closer tolerances. With closed molding it is possible to produce better parts, in less time, with less waste and greatly reduced emissions.

Resin Transfer Molding (RTM)

Resin transfer molding is a process where resin is injected into a closed-cavity mold that is filled with fiber reinforcement (referred to as a preform). It is a relatively low-pressure vacuum (100 psi) process that molds near complete shapes in 30–60 min. It results in two good finished sides, providing the highest quality surface achievable.

The vacuum infusion process is a superior method for construction of composite parts. Parts produced using this method are stronger, lighter, and cheaper to produce. In addition, quality control and quality assurance issues are much easier to deal with than with hand-laid parts. Inspections can be easily carried out before the resin is introduced into the part, and with the use of clear gel coat, the part is very easily examined for flaws after it has been infused. The process is very environmentally friendly. Volatile organic compounds (VOCs) and hazardous air pollutants (HAPs) are drastically reduced. This also means that the working environment is greatly improved. VIP also allows unlimited set-up time because the resin is not catalyzed until all the materials are in place.

Compression Molding

Compression molding is a standard molding technique where a weighed charge is placed into a heated, matched die mold and subjected to mechanically applied pressure. The artifact can be removed after a defined curing schedule (fusion and curing in case of thermoplastic). It is cost-effective process for higher runs. It uses forged steel dies capable of turning out up to 200,000 finished parts. Faster cycle times and lower unit cost are possible when sheet-molding compounds (SMC) are used in the process. Natural fibers including jute, sisal, and kenaf are used as discrete reinforcement. Water, glycerols, glycols, and chitosan solution are used as plasticizers.

Key Applications

Applications in Textiles and Non-wovens

Small-scale and large-scale industries need to produce mass commodities at low cost. Hence, scientific research

on new processing methods and biotechnological engineering aim to develop useful products from natural fibers.

Natural fibers as biocomposites include a wide range of products and different applications ranging from construction or insulation panels made of wood pieces, particles, and fibers, to special textiles (geo-textiles and non-woven textiles), to plastic products based on polymers filled with lignocellulosic particles. Fiber plants are seen as promising lignocellulosic raw materials for different applications. These lignocellulosics are biodegradable, recyclable, and, when combined with natural resin, they are as strong as steel yet of lower density. Such biocomposites may be used in motor vehicles, furniture, and machine construction, insulating materials, gardening and agricultural equipment, and biomedical applications such as screws and total hip replacement stems. Fully resorbable composites are used for internal fracture fixation applications. Partially resorbable composites are used for bone replacement, bone cements, and internal fracture fixation application. Non-resorbable biocomposites are used for spinal fusion in implants, hip or knee joint prostheses, prosthetic sockets, bone plates, dental posts, external fixators, orthodontic brackets, and orthodontic arch-wires.

Examples of applications of composite materials containing lignocellulosic components include glue-lam wood, plywood, plastic board, fire board, medium density fiber, oriented strand board, lignocellulosic mineral particle boards and composites, special functional (water, fire, and bio-resistant), thermosetting polymer composites, thermoplastic polymer composites, natural polymer composites (based on starch, polyhydroxy butyric, and polylactic acid), non-woven, geotextiles, and adsorption chemotextiles (bontonite, carbon active, silica, hydrogel, linoleum).

The production of natural-fiber composites is based on low investments unlike glass fibers. Importing glass fibers can be avoided in Southeast Asian countries where natural fibers are grown abundantly. With rapid advancements in biotechnology, material science, and related fields, the potential for biocomposites is large. Bio-based resources are renewable, widely distributed, available locally, moldable, anisotropic, hydroscopic, recyclable, versatile, nonabrasive, porous, viscoelastic, easily available in many forms, biodegradable, combustible, and reactive. Apart from well-known methods such as dumping, incineration, and recycling, the so-called “bio-composite” could offer a new way for disposal of industrial wastes known as biological decomposition.



Sustainable Technology for Tribological Textiles, Fig. 4
Three layer (sandwich type) nonwoven composite – ramie/kenaf/polypropylene



Sustainable Technology for Tribological Textiles, Fig. 5 Car door panels with biofibers

Fiber from various plants can be used to produce a wide variety of composites. These composites range from value-added specialty products to very large volume commercial materials. High-performance composite materials with uniform densities, durability in adverse environments, and high strength can be produced using agro-based fiber, high-performance adhesives, and fiber modification to overcome dimensional instability, biodegradability, flammability, and degradation caused by ultraviolet light, acids, and bases. Products with complex shapes can also be produced using flexible fiber mats, which can be made by non-woven needling or thermoplastic fiber melt matrix technologies (Fig. 4). Taking advantage of fiber cell wall modification chemistry and combining bast fiber with other materials provides a strategy for producing advanced composites and

Sustainable Technology for Tribological Textiles, Table 1 Fiber type and its applications in hospitals

Fiber type	Fabric structure	Applications
Cotton, viscose, elastomeric-fibre yarns	Woven, nonwoven, knitted	Simple non-elastic and elastic bandages
Alginate fibre, chitosan, silk, viscose, cotton	Woven, nonwoven, knitted	Wound-contact layer
Cotton, viscose, alginate fibre, chitosan	Woven, nonwoven, knitted	Gauze dressing

materials that take advantage of the enhanced properties of all types of materials, and it allows scientists to design materials based on end-use requirements within the framework of cost, availability, renewability, recyclability, sustainability, energy use, and environmental considerations.

Technical Textiles

The natural fibers predominantly used in technical textiles include cotton, jute, silk, and coir. Technical textiles include textiles for automotive applications, medical textiles (e.g., implants), geotextiles (reinforcement of embankments), and protective clothing (e.g., heat and radiation protection for firefighter uniforms, molten metal protection for welders, stab protection, and spacesuits).

When compared with glass fibers traditionally used within the automotive sector, natural fibers demonstrate considerable benefits. The important aspect here is that these benefits are achieved without a significant change in functional performance, such as degradation of mechanical or acoustic properties. The most important benefits include:

- Reduction in cost: Typically, natural fibers are between 25% and 50% cheaper than glass fiber composites.
- Reduction in weight: Four injection-molded ABS car door panels weigh 9 kg. The same panels made using natural fibers weigh only 5 kg and possess equivalent mechanical properties (Fig. 5).
- Lighter weight (less than half the density of glass fibers): This is also an important environmental benefit, as ~85% of energy usage is when a vehicle is being driven. Plant fibers have a maximum density of 1.5 g/cm³ (that of cellulose), which results in high specific strength and stiffness and hence low component weight.
- Better safety: In accident tests, safer crash behavior is observed, particularly in relation to high stability and an absence of splintering, (Marsh 2003).

Medical textiles include all textile goods that find their use in healthcare applications in consumer and medical

usage segments. Application of textiles now goes beyond the usual wound care, incontinence pads, and plasters. The latest innovations include a wide variety of woven, non-woven, and knitted forms of textiles that are increasingly finding their way into a variety of surgical procedures (Table 1). More than 70% of medical textiles are non-woven; this increased share is due to performance, convenience, and disposability. Product innovation by leading brands across the globe has ensured better usage and growth (Urbanczyk 1997; Zhai and Lee 1989).

Antimicrobial agents are also used in the textile sector, principally for hygiene applications. There are several commercial agents that can render a textile antimicrobial. The efficacy of antimicrobial textiles has been proved by standard test methods. Antimicrobial testing and modified Hohenstein tests confirm the viability of antimicrobial finishing. It is suggested that the more specific materials such as secondary metabolites in the selected biomaterials are elegant and sometimes specific in “cidal/static” effects and could be widely used as antimicrobial agents for textile materials.

Engineered textiles combine fabric with glass, ceramics, metal or carbon to produce lightweight hybrids with incredible properties. Sophisticated finishes, such as silicone coatings and holographic laminates, transform color, texture and even form. Smart textiles are no longer a science-fiction fantasy - anti-bacterial, perfume-releasing, self-cleaning and anti-insomniac microfibers are being developed. Stripes of silicone coating speed swimmers through the water; when they come out of it, ceramic fibers utilize solar power to keep them warm (Braddock and Mahony 2005; Mohanty et al. 2002).

References

- S.E. Braddock Clarke, M. O' Mahony, *Techno Textiles: Revolutionary Fabrics for Fashion and Design* (Thames and Hudson, London, 2005), pp. 72–74
- B.J. Collier, P.G. Tortora, *Understanding Textiles*, 6th edn. (Prentice Hall, London, 2001)
- T.F. Cooke, Biodegradability of polymers and fibers – a review of the literature. *J. Polym.* **9**, 171–211 (1990)
- L.T. Drzal, A. Mohanti, M. Mishra, *Natural fibers, Biopolymers and Biocomposites*, CRC Press (2005)

- J.W.S. Hearle, *High-Performance Fibres* (Woodhead Publishing, Cambridge, 2001), pp. 224–231
- W. Liese, Anatomy and properties of bamboo, in *Proceedings of the International Bamboo Workshop*, Hangzhou, 1985, pp. 196–208
- G. Marsh, Next step for automotive materials. *Mater. Today* **6**, 36–43 (2003)
- A.K. Mohanty, M. Mishra, L.T. Drzal, Sustainable bio-composites from renewable resources: opportunities and challenges in the green materials world. *J. Polym. Environ.* **10**(1–2), 19–26 (2002)
- P.N. Prasad, J.E. Mark, T.J. Fai (eds.), *Polymers and other advanced materials: emerging technologies and business opportunities*, in *Proceedings of the 3rd International Conference on Frontiers of Polymers and Advanced Materials*, Kuala Lumpur, 16–20 Jan 1995 (Plenum Press, New York, 1995), pp. 659–665
- O. Romanoschi, S. Romanoschi, J.R. Collier, B.J. Collier, Kenaf alkali processing. *Cellul. Chem. Technol.* **31**(5–6), 347–359 (1997)
- R.M. Rowell, J.S. Han, J.S. Rowell, Characterization and factors effecting fiber properties, in *Natural Polymers and Agro Fibers Composites: Preparation, Properties and Applications*, ed. by F. Elisabete, L.L. Alcides, H.C. Mattso (Emrapa Intrumentaco Agropecuaria, Brasil, 2000), pp. 115–134
- N.D. Saheb, J.P. Jog, Natural fiber polymer composites: a review. *Adv. Polym. Technol.* **18**, 351–363 (1999)
- M. Senthilkumar, J.C. Sakthivel, R. Murugan, Textile composites: an overview. *Indian Text. J.* **117**(9), 93–97 (2007)
- P.M. Smith, *Date Production and Protection* (Food and Agriculture Organization, Rome, 1982)
- G.W. Urbanczyk, in *Applications of Chitin and Chitosan*, ed. by M.F.A. Gooden (Technomic, Lancaster, 1997), p. 281
- C.L. Webber, Crude protein and yield components of six kenaf cultivars as affected by crop maturity. *Ind. Crop. Prod.* **2**, 27 (1993)
- H.M. Zhai, Z.Z. Lee, Ultrastucture of delignification in alkaline pulping of wheat straw. *J. Wood Chem. Technol.* **9**, 387–406 (1989)

Synovial Fluid

- [Polymers in Biotribology](#)

Synovial Lubrication

- [Brush and Hydration Lubrication \(Natural Synovial Joints\)](#)

Synthetic Ester Lubricant Base-Fluids

- [Organic Esters](#)

Synthetic Gear Oils

- [Gear Lubrication](#)

Synthetic Gear Oils and Efficiency

- [Gear Efficiency](#)

System Oils for Large Marine Diesel Engines

- [Marine Engine Oils](#)

